SEMIPARAMETRIC ESTIMATION OF MUTUAL INFORMATION AND RELATED CRITERIA : OPTIMAL TEST OF INDEPENDENCE

AMOR KEZIOU[1] AND PHILIPPE REGNAULT[2]

Abstract. We derive independence tests by means of dependence measures thresholding in a semiparametric context. Precisely, estimates of φ-mutual informations, associated to φ-divergences between a joint distribution and the product distribution of its margins, are derived through the dual representation of φ-divergences. The asymptotic properties of the proposed estimates are established, including consistency, asymptotic distributions and large deviations principle. The obtained tests of independence are compared via their relative asymptotic Bahadur efficiency and numerical simulations. It follows that the proposed semiparametric Kullback-Leibler Mutual information test is the optimal one. On the other hand, the proposed approach provides a new method for estimating the Kullback-Leibler mutual information in a semiparametric setting, as well as a model selection procedure in large class of dependency models including semiparametric copulas.

Keywords : Mutual informations, φ-divergences, Fenchel Duality, Tests of independence, semiparametric inference.

Contents

Date: July 2015.

1. Introduction and notations

Measuring the dependence between random variables has been a central aim of probability theory since its earliest developments. Classical examples of dependence measures are correlation measures of Pearson, Kendall or Spearman. While the first one focuses on linear relationship between real random variables, the two second ones measure the monotonic relationship between variables taking values in ordered sets. Pure-independence measures, between variables X and Y taking values in general measurable spaces $(\mathcal{X}, \mathcal{A}_{\mathcal{X}})$ and $(\mathcal{Y}, \mathcal{A}_{\mathcal{Y}})$, can be defined by considering any divergence between the joint distribution $\mathbb{P}$ of (X, Y) and the product distribution of its margins $\mathbb{P}^{\perp} := \mathbb{P}_1 \otimes \mathbb{P}_2$, where $\mathbb{P}_1$ and $\mathbb{P}_2$ are, respectively, the marginal distributions of X and Y. The most outstanding and widely used example of such dependence measures is the χ^2-divergence between $\mathbb{P}$ and $\mathbb{P}^{\perp}$ defined by

$$\chi^2(\mathbb{P}, \mathbb{P}^{\perp}) := \frac{1}{2} \int_{\mathcal{X}\times\mathcal{Y}} \left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{P}^{\perp}}(x, y) - 1 \right)^2 \mathrm{d}\mathbb{P}^{\perp}(x, y), \tag{1}$$

where $\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{P}^{\perp}}$ denotes the density of $\mathbb{P}$ with respect to (w.r.t.) $\mathbb{P}^{\perp}$. Note that, if $\mathbb{P}$ is a discrete distribution, i.e., if its support $\mathcal{X} \times \mathcal{Y} := \mathrm{supp}(\mathbb{P})$ is discrete (finite or countably infinite) set, then the above divergence writes

$$\chi^2(\mathbb{P}, \mathbb{P}^{\perp}) = \frac{1}{2} \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \frac{(p_{x,y} - p_x p_y)^2}{p_x p_y},$$

where $\mathbb{P} := (p_{x,y})_{(x,y)}$, $\mathbb{P}^{\perp} = (p_x p_y)_{(x,y)}$, with $p_x := \sum_y p_{x,y}$ and $p_y := \sum_x p_{x,y}$. Another classical example, associated to the Kullback-Leibler (KL) divergence between $\mathbb{P}$ and $\mathbb{P}^{\perp}$, is the well-known mutual information (MI) defined by (see e.g. Cover and Thomas (2006))

$$I_{KL}(\mathbb{P}) := \mathbb{K}(\mathbb{P}, \mathbb{P}^{\perp}) := \int_{\mathcal{X}\times\mathcal{Y}} \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{P}^{\perp}}(x, y) \, \log \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{P}^{\perp}}(x, y) \, \mathrm{d}\mathbb{P}^{\perp}(x, y), \tag{2}$$

which, in the case of discrete distributions, can be written under the form

$$I_{KL}(\mathbb{P}) = \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p_{x,y} \log \frac{p_{x,y}}{p_x p_y}.$$

We will call the above classical measures of dependence (1) and (2), respectively, χ^2-mutual information (χ^2-MI) and KL-mutual information (KL-MI). When dealing with i.i.d. observations $(X_1, Y_1), \ldots, (X_n, Y_n)$, of two random variables (X, Y), we may test the null hypothesis, that the variables X and Y are independent, by means of estimating such dependence measure and deciding to reject the null hypothesis of independence if the estimate is sufficiently far from zero; the classical χ^2-independence test is such a procedure : the corresponding test statistic (in the discrete-distribution case) is

$$2n\,\chi^2\left(\widehat{\mathbb{P}}, \widehat{\mathbb{P}}^{\perp}\right) = n \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \frac{(\widehat{p}_{x,y} - \widehat{p}_x\widehat{p}_y)^2}{\widehat{p}_x\widehat{p}_y}, \tag{3}$$

where $\widehat{\mathbb{P}} := (\widehat{p}_{x,y})_{(x,y)}$ and $\widehat{\mathbb{P}}^{\perp} := (\widehat{p}_x\,\widehat{p}_y)_{(x,y)}$ are, respectively, the empirical versions of $\mathbb{P} = (p_{x,y})_{(x,y)}$ and $\mathbb{P}^{\perp} = (p_x p_y)_{(x,y)}$. Likewise, to test the independence, we can consider as dependence measure the KL-MI and use the test statistic

$$2n\,I_{KL}(\widehat{\mathbb{P}}) = 2n \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \widehat{p}_{x,y} \log \frac{\widehat{p}_{x,y}}{\widehat{p}_x\widehat{p}_y}. \tag{4}$$

The dependence measure can also be any other φ-divergence between $\mathbb{P}$ and $\mathbb{P}^{\perp}$. The tests based on such dependence measures, including the χ^2-MI and KL-MI ones, have been extensively studied in the case of finite-discrete distributions; see e.g. Pardo (2006) Chapter 8, and the references therein. When dealing with continuous distributions (or continuous random variables), obviously, the above direct plug-in estimates (3) and (4), of the dependence measures (1) and (2), are not well defined. Moreover, for countably-infinite discrete distributions, although the above estimates (3) and (4) remain well defined, their limiting distributions are not accessible. Therefore, in the case of non finite-discrete distributions, particularly, for the widely used KL-MI, other kind of estimates have been proposed and studied in the literature; see e.g. Moon *et al.* (1995) for a kernel density estimate, Kraskov *et al.* (2004) for a k-nearest-neighbor estimate extending those of Shannon entropy in one dimension based on m-spacing; see e.g. Tsybakov and van der Meulen (1996), Dudewicz and van der Meulen (1981) and Beirlant *et al.* (1997) among others. Van Hulle (2005) derive an estimate using Edgeworth approximation of Shannon entropy. Darbellay and Vajda (1999), Wang *et al.* (2005) and Cellucci *et al.* (2005) propose estimates based on adaptative partitioning of $\mathcal{X} \times \mathcal{Y}$. See also Khan *et al.* (2007) for an overview and numerical comparisons of these estimates. Based on the Kullback-Leibler importance estimation procedure, see Sugiyama *et al.* (2008), Suzuki *et al.* (2008) obtain an

estimate of KL-MI called maximum likelihood mutual information, see also Sugiyama *et al.* (2012) Chapter 11. Unfortunately, their (asymptotic) distributions remain inaccessible. Hence, testing independence from these estimates requires Monte-Carlo or Bootstrap approximations of the related p-values. On the other hand, the above nonparametric estimates suffer from loss of efficiency, due to smoothing or partitioning, and suffer also from the difficulty of conveniently choosing the classes, the number of classes or the smoothing parameters (the bandwidths and the kernels). The present paper introduces new efficient semiparametric estimates of φ-mutual information (φ-MI), i.e., dependence measures associated to φ-divergence functionals, including the well known KL-MI and χ^2-MI. These estimates are obtained by making use of a dual representation of φ-MI, presented in Section 2, without using any smoothing nor partitioning. The obtained estimates are defined in the same way for both finite-discrete or non-discrete distributions, and coincide with the direct plug-in ones in the case of finite-discrete distributions. Their asymptotic properties are presented in Section 3. Particularly, the consistency is stated for a large variety of semiparametric models for $d\mathbb{P}/d\mathbb{P}^{\perp}$; the asymptotic distribution is obtained for the KL-MI estimate in a special setting. The present approach leads to new independence tests, whose Bahadur efficiency are compared in Section 4 ; the most efficient test is shown to be the one based on the proposed estimate of the particular KL-MI criterion. It can be used also in order to build a large variety of dependence models, through for instance a cross validation-type model selection procedure based on the proposed estimate of φ-MI measure of dependence; see Section 2.4. The powers of φ-MI based tests are compared numerically to classical noncorrelation tests in Section 5. The results in the present paper have the advantage (unlike the classical noncorrelation tests) to remain valid in the case of multisample problem (estimating φ-mutual informations of a multidimensional random variable as well as testing simultaneous independence of its components), but for simplicity, the results will be presented only for the two-sample case. The same results hold for the multisample problem. All proofs are postponed to the Appendix.

2. φ-mutual informations, Dual representations and Estimation strategy

Given an i.i.d. sample, $(X_1, Y_1), \ldots, (X_n, Y_n)$, of a random vector (X, Y) taking values in a measurable space $(\mathcal{X} \times \mathcal{Y}, \mathcal{A}_{\mathcal{X}} \otimes \mathcal{A}_{\mathcal{Y}})$, we aim at testing the null hypothesis $\mathcal{H}_0$ of independence of the margins X and Y; formally

$$\mathcal{H}_0 : X \text{ and } Y \text{ are independent}, \quad \text{against} \quad \mathcal{H}_1 : X \text{ and } Y \text{ are dependent}. \tag{5}$$

We derive such tests by estimating and thresholding φ-mutual informations between X and Y in a semiparametric context. Sections 2.1, 2.2 and 2.3 to follow, respectively, define φ-mutual informations, present the semiparametric model under study, and introduce estimates

of φ-MI used as test statistics for the test problem (5). Section 2.4 defines a cross-validation procedure for model selection among L candidate models for the ratio $\mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{P}^{\perp}$, using the proposed estimate of φ-MI.

2.1. **Introducing φ-mutual informations.** Denote by $\mathcal{M}_1(\mathcal{X}\times\mathcal{Y})$ the set of all probability distributions on the product measurable space $(\mathcal{X}\times\mathcal{Y}, \mathcal{A}_{\mathcal{X}}\otimes\mathcal{A}_{\mathcal{Y}})$. Let $\varphi : \mathbb{R}\to[0,+\infty]$ be some nonnegative closed proper convex function such that its domain $\mathrm{dom}_{\varphi} := \{x\in\mathbb{R}; \varphi(x)<\infty\} =: (a_{\varphi}, b_{\varphi})$ is an interval, with endpoints $a_{\varphi}<1<b_{\varphi}$, and $\varphi(1)=0$. The interval $(a_{\varphi}, b_{\varphi})$ may be bounded or unbounded, open or not. The φ-divergence between any probability distributions $Q, P\in\mathcal{M}_1(\mathcal{X}\times\mathcal{Y})$, if Q is absolutely continuous with respect to (a.c.w.r.t.) P, is defined by

$$D_{\varphi}(Q,P) := \int_{\mathcal{X}\times\mathcal{Y}} \varphi\left(\frac{dQ}{dP}(x,y)\right) dP(x,y).$$

If Q is not a.c.w.r.t. P, we set $D_{\varphi}(Q,P)=+\infty$. Note that $D_{\varphi}(Q,P)\geq 0$, for any Q and P. Moreover, if φ is strictly convex on some neighborhood of 1, we have the fundamental property

$$D_{\varphi}(Q,P)\geq 0, \text{ with equality if and only if } Q=P.$$

In the following, we assume that the function φ is strictly convex and two times continuously differentiable on the interior of its domain $(a_{\varphi}, b_{\varphi})$. We have then $\varphi'(1)=0$, and without loss of generality, we can assume that $\varphi''(1)=1$. The well-known Kullback-Leibler divergence $\mathbb{K}(\cdot,\cdot)$ is obtained for $\varphi(x)=\varphi_1(x) := x\log x - x + 1$, the "modified" Kullback-Leibler divergence $\mathbb{K}_m(\cdot,\cdot)$ is obtained for $\varphi(x)=\varphi_0(x) := -\log x + x - 1$. The χ^2 and modified-χ^2 divergences, denoted $\chi^2(\cdot,\cdot)$ and $\chi^2_m(\cdot,\cdot)$, are associated, respectively, to the convex functions $\varphi(x)=\varphi_2(x) := (x-1)^2/2$ and $\varphi(x)=\varphi_{-1}(x) := (x-1)^2/(2x)$. The so-called Hellinger distance $H(\cdot,\cdot)$ is obtained for $\varphi(x)=\varphi_{1/2}(x) := 2(\sqrt{x}-1)^2$; see Table 1. All these divergences are members of the so-called "power-divergences" $D_{\varphi_\gamma}(\cdot,\cdot)$ associated to the convex functions $\varphi_\gamma(\cdot)$ defined by

$$\varphi_\gamma(\cdot) : x\in\mathbb{R}_+^* \mapsto \varphi_\gamma(x) := \frac{x^\gamma - \gamma x + \gamma - 1}{\gamma(\gamma-1)} \tag{6}$$

if $\gamma\in\mathbb{R}\setminus\{0,1\}$, $\varphi_0(x) := -\log x + x - 1$ and $\varphi_1(x) := x\log x - x + 1$. The standard divergences $\mathbb{K}(\cdot,\cdot)$, $\mathbb{K}_m(\cdot,\cdot)$, $\chi^2(\cdot,\cdot)$, $\chi^2_m(\cdot,\cdot)$ and $H(\cdot,\cdot)$ are then associated, respectively, to the real convex functions $\varphi_1(\cdot)$, $\varphi_0(\cdot)$, $\varphi_2(\cdot)$, $\varphi_{-1}(\cdot)$ and $\varphi_{1/2}(\cdot)$. Note that the divergences are generally not symmetric; particularly, we have for any $Q, P\in\mathcal{M}_1(\mathcal{X}\times\mathcal{Y})$, $\mathbb{K}_m(Q,P)=\mathbb{K}(P,Q)$ and $\chi^2_m(Q,P)=\chi^2(P,Q)$. For more details and proofs, we can refer to Liese and Vajda (1987) and Broniatowski and Keziou (2006). For any probability distribution $P\in\mathcal{M}_1(\mathcal{X}\times\mathcal{Y})$, let $P^{\perp}$ denotes the product distribution $P^{\perp} := P_1\otimes P_2$ of the margins P_1 and P_2 of P. The φ-mutual

information of P, associated to the divergence $D_\varphi(\cdot,\cdot)$, is defined as

$$I_\varphi(P) := D_\varphi(P, P^\perp).$$

For any random vector (X, Y) defined on a probability space $(\Omega, \mathcal{A}, \mathbf{P})$ and taking its values in $(\mathcal{X} \times \mathcal{Y}, \mathcal{A}_\mathcal{X} \otimes \mathcal{A}_\mathcal{Y})$, with joint distribution $\mathbb{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$, the φ-mutual information (φ-MI) of (X, Y) is defined to be

$$I_\varphi(X, Y) := I_\varphi(\mathbb{P}) = D_\varphi(\mathbb{P}, \mathbb{P}^\perp) = \int_{\mathcal{X}\times\mathcal{Y}} \varphi\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{P}^\perp}(x, y)\right) \mathrm{d}\mathbb{P}^\perp(x, y). \tag{7}$$

Since $D_\varphi(\mathbb{P}, \mathbb{P}^\perp) \geq 0$, with equality if and only if $\mathbb{P} = \mathbb{P}^\perp$, i.e., if and only if X and Y are independent, φ-MI measures then the dependence between the random variables X and Y. In contrast to the correlation coefficients of Pearson, Kendall or Spearman, the φ-MI does not focus on the linear or monotonic relationship between random variables; it constitutes a proper dependency measure. Note that I_{φ_1} and I_{φ_2}, with φ_1 and φ_2 given in Table 1, are, respectively, the KL-MI and χ^2-MI, given by (2) and (1). Thus, the test problem (5) is equivalent, in the context of I_φ criteria, to testing

$$I_\varphi(\mathbb{P}) = 0 \quad \text{against} \quad I_\varphi(\mathbb{P}) > 0.$$

Hence, we can use as test statistic an estimate of $I_\varphi(\mathbb{P})$, and reject the null hypothesis $\mathcal{H}_0$ when the estimate takes large values. A natural attempt to estimate the φ-MI of (X, Y) consists in considering the plug-in estimate of $I_\varphi(\mathbb{P})$ obtained by replacing $\mathbb{P}(\cdot)$ by its empirical counterpart

$$\widehat{\mathbb{P}}(\cdot) = \frac{1}{n}\sum_{i=1}^{n} \delta_{(X_i,Y_i)}(\cdot), \tag{8}$$

associated to the i.i.d. sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of (X, Y). Here, $\delta_{(x,y)}(\cdot)$ denotes the Dirac measure at (x, y) for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Unfortunately, by doing so, we only measure dependence of the contingency table associated to the sample. When dealing with variables X and Y absolutely continuous with respect to Lebesgue measure, the contingency table is almost surely an $n \times n$ table with all coefficients except diagonal ones equal to zero ; particularly, variables X and Y appear (misleadingly) purely dependent, yielding to reject systematically the null hypothesis. A second, less crude, approach consists in gathering the values X_i and Y_i into classes and testing independence between the induced finite-discrete variables $\tilde{X}$ and $\tilde{Y}$, by empirically estimating the φ-MI of $(\tilde{X}, \tilde{Y})$. This widespread approach suffers from the difficulty of conveniently choosing the classes. Moreover, an important amount of information carried by the sample is lost during this process, yielding to poor efficiency – or power – of these tests. An other approach, is to use kernel nonparametric estimates of the joint density and the marginal ones, but as it is well known this provides less efficient estimates and leads

to the difficulty of choosing the optimal smoothing parameters. As an alternative, we propose in the present paper semiparametric modeling of the ratio $\mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{P}^{\perp}$, and the use of duality to obtain well-defined estimates of φ-MI without smoothing nor partitioning. The present approach applies for both continuous or discrete distributions, or mixtures of continuous and discrete distirubtions.

2.2. **Semiparametric modeling of the ratio $\mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{P}^{\perp}$.** Assume that the joint distribution $\mathbb{P}$ of the random vector (X, Y) belongs to the semiparametric model

$$\mathcal{M}_{\Theta} := \left\{ P \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y}) \text{ such that } \frac{dP}{dP^{\perp}}(\cdot,\cdot) =: h_{\theta}(\cdot,\cdot);\, \theta \in \Theta \right\}, \tag{9}$$

where $\Theta \subset \mathbb{R}^{1+d}$ is the parameter space, and $h_{\theta}(\cdot,\cdot) : (x, y) \in \mathcal{X} \times \mathcal{Y} \mapsto h_{\theta}(x, y) \in \mathbb{R}$ is some specified real-valued function, indexed by the parameter θ. In the sequel, we will consider the following assumptions on the model $\mathcal{M}_{\Theta}$.

(A.1) $(h_{\theta}(x, y) = h_{\theta'}(x, y), \forall (x, y) \in \mathcal{X} \times \mathcal{Y}) \;\Rightarrow\; (\theta = \theta')$ (identifiability);
(A.2) there exists (a unique) $\theta_0 \in \mathrm{int}(\Theta)$ satisfying $h_{\theta_0}(x, y) = 1$, $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}$.

Assumption (A.1) is a natural identifiability condition for $dP/dP^{\perp}$. Assumption (A.2) ensures independence is covered by the model $\mathcal{M}_{\Theta}$. The uniqueness of θ_0 follows from Assumption (A.1). Denote by θ_T the "true" unknown value of the parameter, namely, the unique value satisfying

$$\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{P}^{\perp}}(x, y) = h_{\theta_T}(x, y), \quad \forall (x, y) \in \mathcal{X} \times \mathcal{Y},$$

which is assumed to be an interior point of Θ. Then, we have $\theta_T = \theta_0$ if and only if X and Y are independent. Below are listed some relevant examples of the model (9).

Example 2.1. *Let $(X, Y) \in \mathbb{R}^2$ be a centered Gaussian random vector with correlation coefficient $\rho \in]-1, 1[$ and centered normal margins with the same variance $\sigma^2 > 0$. A straightforward computation shows that the ratio $\mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{P}^{\perp}$ can be written under the form of the model (9) where*

$$h_{\theta}(x, y) = \exp\left\{\alpha + \beta_1(x^2 + y^2) + \beta_2 xy\right\}, \tag{10}$$

$\theta := (\alpha, \beta_1, \beta_2)^{\top} \in \mathbb{R}^3$, with $\alpha = -\log(1 - \rho^2)/2$, $\beta_1 = -\rho^2/(2\sigma^2(1 - \rho^2))$ and $\beta_2 = \rho/(\sigma^2(1 - \rho^2))$. Note that the parameter value, corresponding to the independence hypothesis, is $\theta_0 = (0, 0, 0)^{\top}$. Moreover, if the distribution of (X, Y) is Gaussian with unknown mean $\mu := (\mu_1, \mu_2)\top$ and unknown variance matrix Γ, then we can show that the ratio $\mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{P}^{\perp}$ can be written under the form of the model (9) with

$$h_{\theta}(x, y) = \exp\left\{\alpha + \beta_1 x + \beta_2 y + \beta_3 x^2 + \beta_4 y^2 + \beta_5 xy\right\}, \tag{11}$$

and $\theta := (\alpha, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)^\top$. *Note that the number of free parameters in* θ_T *is* $d = 5$, *and that* α_T *is considered as a normalizing parameter due to the constraint* $\int_{\mathcal{X}\times\mathcal{Y}} h_{\theta_T}(x,y)\,\mathrm{d}\mathbb{P}^\perp(x,y) = \int_{\mathcal{X}\times\mathcal{Y}} \mathrm{d}\mathbb{P}(x,y) = 1$ *since* $\mathbb{P}$ *is a probability distribution. Moreover, we have* $\theta_0 = (0,\ldots,0)^\top \in \mathbb{R}^6$.

Example 2.2. *Let* $\psi_0(\cdot,\cdot) := \mathbb{1}_{\mathcal{X}\times\mathcal{Y}}(\cdot,\cdot), \psi_1(\cdot,\cdot), \psi_2(\cdot,\cdot), \ldots,$ *be some basis functions of the space* $L^2(\mathcal{X}\times\mathcal{Y}, \mathbb{P}^\perp)$, *and assume that* $\log(\mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{P}^\perp(\cdot,\cdot)) \in L^2\left(\mathcal{X}\times\mathcal{Y}, \mathbb{P}^\perp\right)$. *We can then build increasing models of the form (9) developing the function*

$$(x,y) \in \mathcal{X}\times\mathcal{Y} \mapsto \log \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{P}^\perp}(x,y)$$

according to the above basis functions. Using for instance the first $(1+d)$*-basis functions, we obtain the following model for* $\mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{P}^\perp(\cdot,\cdot)$

$$h_\theta : (x,y) \in \mathcal{X}\times\mathcal{Y} \mapsto h_\theta(x,y) = \exp\left(\alpha + \beta_1\psi_1(x,y) + \cdots + \beta_d\psi_d(x,y)\right),$$

where $\theta = (\alpha, \beta_1, \ldots, \beta_d)^\top \in \Theta \subset \mathbb{R}^{1+d}$. Then, the independence parameter value is $\theta_0 = (0,\ldots,0)^\top \in \mathbb{R}^{1+d}$.

Example 2.3. *Assume that the support of* $\mathbb{P}$, $supp(\mathbb{P}) =: \mathcal{X}\times\mathcal{Y}$, *is a known finite-discrete set of size* K_1K_2; *denote by* $(\mathbb{P}(x,y))_{(x,y)\in\mathcal{X}\times\mathcal{Y}} := (p_{x,y})_{(x,y)\in\mathcal{X}\times\mathcal{Y}}$ *the density of* $\mathbb{P}$ *with respect to the counting measure on* $\mathcal{X}\times\mathcal{Y}$. *Then we have*

$$\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{P}^\perp}(x,y) = \exp\left(\sum_{(a,b)\in\mathcal{X}\times\mathcal{Y}} \theta_{a,b}\ \mathbb{1}_{\{a\}}(x)\,\mathbb{1}_{\{b\}}(y)\right), \tag{12}$$

where

$$\theta_{a,b} = \log\frac{p_{a,b}}{p_a p_b}, \quad (a,b) \in \mathcal{X}\times\mathcal{Y}.$$

If we denote for instance the elements of $\mathcal{X}$ *and* $\mathcal{Y}$ *as follows*

$$\mathcal{X} := \{a_1,\ldots,a_{K_1}\} \quad \textit{and} \quad \mathcal{Y} := \{b_1,\ldots,b_{K_2}\},$$

then we can see that $\mathbb{P}$ *belongs to the model (9) taking*

$$h_\theta(x,y) = \exp\left(\alpha + \sum_{(i,j)\neq(1,1)} \beta_{i,j}\,\mathbb{1}_{\{a_i\}}(x)\,\mathbb{1}_{\{b_j\}}(y)\right), \tag{13}$$

with the parametrization $\theta = (\alpha, \beta^\top)^\top \in \mathbb{R}^{K_1K_2}$, *where* α *is a scalar and* $\beta = (\beta_{i,j})_{(i,j)\neq(1,1)}$ *is the* (K_1K_2-1)*-dimensional vector obtained from the* $K_1\times K_2$*-matrix of real entries* $(\beta_{i,j})$ *removing the first entry* $\beta_{1,1}$. *Moreover, we have for the true value* θ_T

$$\alpha_T = \log\frac{p_{a_1,b_1}}{p_{a_1}p_{b_1}}, \quad \textit{and} \quad \beta_{i,j_T} = \log\frac{p_{a_i,b_j}}{p_{a_i}p_{b_j}} - \log\frac{p_{a_1,b_1}}{p_{a_1}p_{b_1}},$$

for all $(i,j) \in \{1,\dots,K_1\} \times \{1,\dots,K_2\} \setminus \{(1,1)\}$, *and that the number of free-parameters in* θ_T *is equal to* $(K_1-1)(K_2-1)$. *Moreover, we have* $\theta_0 = (0,\dots,0)^\top \in \mathbb{R}^{K_1K_2}$.

Example 2.4. *Assume that the distribution* $\mathbb{P}$ *of the random vector* $(X,Y) \in \mathbb{R}^2$ *is of continuous margins. The copula* $C(\cdot,\cdot)$ *of the vector* (X,Y), *see e.g. Nelsen (2006), is defined,* $\forall (u,v) \in]0,1[^2$, *by*

$$C(u,v) := F(F_1^{-1}(u), F_2^{-1}(v)),$$

where $F(\cdot,\cdot)$ *is the cumulative distribution function of the vector* (X,Y), *and* F_1 *and* F_2 *are the (marginal) cumulative distribution functions of* X *and* Y, *respectively. The copula* $C(\cdot,\cdot)$ *is in itself a distribution function on* $]0,1[^2$. *If* $F(\cdot,\cdot)$ *is absolutely continuous with respect to the Lebesgue measure on* $\mathbb{R}^2$, *then we have the relation*

$$\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{P}^\perp}(x,y) = \frac{f(x,y)}{f_1(x)f_2(y)} = c\left(F_1(x), F_2(y)\right),$$

where $f(\cdot,\cdot)$ *is the joint density of* (X,Y), f_1 *and* f_2 *are the marginal densities of* X *and* Y, *and* $c(\cdot,\cdot)$ *the copula density. Numerous parametric examples of the model (9) can then be obtained taking the function*

$$h_\theta(x,y) = c_\beta(F_{1,\gamma_1}(x), F_{2,\gamma_2}(y)) \tag{14}$$

where $\{c_\beta(\cdot,\cdot);\, \beta \in D \subset \mathbb{R}^m\}$ *is some parametric copula density model, see e.g. Nelsen (2006) or Joe (1997) for examples of such models, and* $\{F_{1,\gamma_1}; \gamma_1 \in \Gamma_1\}$ *and* $\{F_{2,\gamma_2}; \gamma_2 \in \Gamma_2\}$ *are some parametric models for the marginal distribution functions. Here, the parameter of interest is* $\theta := (\gamma_1, \gamma_2, \beta) \in \Theta := \Gamma_1 \times \Gamma_2 \times D$. *Note that the assumption (A.2) is generally not satisfied for this particular model. In fact, if we denote* β_0 *the particular value corresponding to the copula of independence, then we have* $h_{(\gamma_1,\gamma_2,\beta_0)}(\cdot,\cdot) = 1$ *for any* $(\gamma_1,\gamma_2) \in \Gamma_1 \times \Gamma_2$. *Although assumption (A.2) is generally not satisfied, models (14) can be used in estimating* φ*-MI under the assumption that the margins are dependent.*

Example 2.5. *We can also deal with semiparametric models induced by semiparametric models of copula densities, with nonparametric unknown continuous marginal distribution functions* $F_1(\cdot)$ *and* $F_2(\cdot)$, *taking*

$$h_\theta(x,y) = c_\theta(F_1(x), F_2(y));\ \theta \in \Theta \subset \mathbb{R}^d.$$

2.3. **Dual representation and dual estimation of φ-MI.** We define estimates of φ-MI by taking advantage of the modeling (9) and the dual representation of φ-divergences obtained in Keziou (2003) and Broniatowski and Keziou (2006). Denote $\varphi^*(\cdot)$ the convex conjugate of the convex function $\varphi(\cdot)$, namely, the function defined by

$$\varphi^* : t \in \mathbb{R} \mapsto \varphi^*(t) := \sup_{x \in \mathbb{R}} \{tx - \varphi(x)\} \in \mathbb{R} \cup \{+\infty\}.$$

Note that $\varphi^*(\cdot)$ is, in turn, a proper closed convex function, in particular, $\varphi^*(0) = 0$. Assume that $\varphi(\cdot)$ is essentially smooth, i.e., differentiable on $]a_\varphi, b_\varphi[$ with $\lim_{x\downarrow a_\varphi} \varphi'(x) = -\infty$ if $a_\varphi > -\infty$ and $\lim_{x\uparrow b_\varphi} \varphi'(x) = +\infty$ if $b_\varphi < +\infty$. This is equivalent to the condition that $\varphi^*(\cdot)$ is strictly convex on its domain. Provided that

(A.3) the φ-mutual information $I_\varphi(\mathbb{P}) < \infty$,

see its definition (7), it can be rewritten under the form

$$I_\varphi(\mathbb{P}) = \sup_{f\in\mathcal{F}} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} f(x,y)\, \mathrm{d}\mathbb{P}(x,y) - \int_{\mathcal{X}\times\mathcal{Y}} \varphi^*\left(f(x,y)\right)\, \mathrm{d}\mathbb{P}^\perp(x,y) \right\}, \tag{15}$$

where $\mathcal{F}$ is any class, of measurable real-valued functions $f : \mathcal{X}\times\mathcal{Y} \to \overline{\mathbb{R}}$, that contains the particular function $\varphi'(\mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{P}^\perp)$ and satisfies the condition $\int_{\mathcal{X}\times\mathcal{Y}} |f|\, \mathrm{d}\mathbb{P} < \infty$, for all $f \in \mathcal{F}$. Note that, for all $x \in (a_\varphi, b_\varphi)$, we have

$$\varphi^*(\varphi'(x)) = x\varphi'(x) - \varphi(x).$$

In Table 1 are given explicit formulas of convex conjugates of some standard divergences. From

$D_\varphi(\cdot,\cdot)$	$\varphi(\cdot)$	$\mathrm{dom}\varphi$	$\mathrm{dom}\varphi^*$	$\varphi^*(\cdot)$
$\mathbb{K}_m(\cdot,\cdot)$	$\varphi_0(x) := -\log x + x - 1$	$]0,+\infty[$	$]-\infty, 1[$	$-\log(1-t)$
$\mathbb{K}(\cdot,\cdot)$	$\varphi_1(x) := x\log x - x + 1$	$[0,+\infty[$	$\mathbb{R}$	$e^t - 1$
$\chi^2_m(\cdot,\cdot)$	$\varphi_{-1}(x) := \frac{1}{2}\frac{(x-1)^2}{x}$	$]0,+\infty[$	$]-\infty, \frac{1}{2}]$	$1 - \sqrt{1-2t}$
$\chi^2(\cdot,\cdot)$	$\varphi_2(x) := \frac{1}{2}(x-1)^2$	$\mathbb{R}$	$\mathbb{R}$	$\frac{1}{2}t^2 + t$
$H(\cdot,\cdot)$	$\varphi_{1/2}(x) := 2(\sqrt{x}-1)^2$	$[0,+\infty[$	$]-\infty, 2[$	$\frac{2t}{2-t}$

TABLE 1. Convex conjugates for some standard divergences.

(15), taking into account the model (9) by specifying

$$\mathcal{F} = \{\varphi'(h_\theta);\, \theta \in \Theta\},$$

and assuming in addition that

(A.4) for all $\theta \in \Theta$, we have $\int_{\mathcal{X}\times\mathcal{Y}} |\varphi'(h_\theta(x,y))|\, \mathrm{d}\mathbb{P}(x,y) < \infty$,

we obtain

$$I_\varphi(\mathbb{P}) = \sup_{\theta\in\Theta} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} \varphi'(h_\theta(x,y))\, \mathrm{d}\mathbb{P}(x,y) - \int_{\mathcal{X}\times\mathcal{Y}} \varphi^*\left(\varphi'(h_\theta(x,y))\right)\, \mathrm{d}\mathbb{P}^\perp(x,y) \right\}. \tag{16}$$

Moreover, the supremum is unique and achieved in $\theta = \theta_T$. The uniqueness of the supremum θ_T follows from the strict convexity of $\varphi^*(\cdot)$ and the identifiability assumption (A.1). We propose

then the following "dual" estimate of $I_\varphi(\mathbb{P})$

$$\begin{aligned}\widehat{I}_\varphi &:= \sup_{\theta\in\Theta}\left\{\int_{\mathcal{X}\times\mathcal{Y}}\varphi'\left(h_\theta(x,y)\right)\,\mathrm{d}\widehat{\mathbb{P}}(x,y)-\int_{\mathcal{X}\times\mathcal{Y}}\varphi^*\left(\varphi'(h_\theta(x,y))\right)\,\mathrm{d}\widehat{\mathbb{P}}_1\otimes\widehat{\mathbb{P}}_2(x,y)\right\}\\ &= \sup_{\theta\in\Theta}\left\{\frac{1}{n}\sum_{i=1}^n\varphi'\left(h_\theta(X_i,Y_i)\right)-\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n\varphi^*\left(\varphi'\left(h_\theta(X_i,Y_j)\right)\right)\right\},\end{aligned}\tag{17}$$

and the following "dual" estimate of the parameter θ_T

$$\begin{aligned}\widehat{\theta}_\varphi &:= \arg\sup_{\theta\in\Theta}\left\{\int_{\mathcal{X}\times\mathcal{Y}}\varphi'\left(h_\theta(x,y)\right)\,\mathrm{d}\widehat{\mathbb{P}}(x,y)-\int_{\mathcal{X}\times\mathcal{Y}}\varphi^*\left(\varphi'(h_\theta(x,y))\right)\,\mathrm{d}\widehat{\mathbb{P}}_1\otimes\widehat{\mathbb{P}}_2(x,y)\right\}\\ &= \arg\sup_{\theta\in\Theta}\left\{\frac{1}{n}\sum_{i=1}^n\varphi'\left(h_\theta(X_i,Y_i)\right)-\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n\varphi^*\left(\varphi'\left(h_\theta(X_i,Y_j)\right)\right)\right\},\end{aligned}\tag{18}$$

where $\widehat{\mathbb{P}}(\cdot)$ is the empirical distribution, associated to the sample, given by (8). For ease of presentation, define, $\forall\theta\in\Theta$ and $\forall(x,y)\in\mathcal{X}\times\mathcal{Y}$, the functions

$$f_\theta(x,y) := \varphi'(h_\theta(x,y)),\tag{19}$$

$$g_\theta(x,y) := \varphi^*\left(\varphi'(h_\theta(x,y))\right) = h_\theta(x,y)\varphi'\left(h_\theta(x,y)\right)-\varphi\left(h_\theta(x,y)\right),\tag{20}$$

which we assume to be continuous, in θ, on the set Θ,

$$M:\theta\in\Theta\mapsto M(\theta) := \int_{\mathcal{X}\times\mathcal{Y}}f_\theta(x,y)\,\mathrm{d}\mathbb{P}(x,y)-\int_{\mathcal{X}\times\mathcal{Y}}g_\theta(x,y)\,\mathrm{d}\mathbb{P}_1\otimes\mathbb{P}_2(x,y)\tag{21}$$

and its empirical version

$$M_n:\theta\in\Theta\mapsto M_n(\theta) := \int_{\mathcal{X}\times\mathcal{Y}}f_\theta(x,y)\,\mathrm{d}\widehat{\mathbb{P}}(x,y)-\int_{\mathcal{X}\times\mathcal{Y}}g_\theta(x,y)\,\mathrm{d}\widehat{\mathbb{P}}_1\otimes\widehat{\mathbb{P}}_2(x,y).\tag{22}$$

Therefore, the formula (16) becomes

$$I_\varphi(\mathbb{P}) = \sup_{\theta\in\Theta}M(\theta) = M(\theta_T),\quad\text{and}\quad\theta_T = \arg\sup_{\theta\in\Theta}M(\theta).\tag{23}$$

The estimates (17) and (18), in turn, can be written as

$$\widehat{I}_\varphi = \sup_{\theta\in\Theta}M_n(\theta) = M_n(\widehat{\theta}_\varphi)\tag{24}$$

and

$$\widehat{\theta}_\varphi = \arg\sup_{\theta\in\Theta}M_n(\theta).\tag{25}$$

Note that the functions $f_\theta(\cdot,\cdot)$, $g_\theta(\cdot,\cdot)$, $M(\cdot)$ and $M_n(\cdot)$ all depend on $\varphi(\cdot)$, but the subscript φ is omitted for simplicity.

Example 2.6. *In the context of finite-discrete distributions, using the exponential model described in Example 2.3, we show that the proposed dual estimate (17) of $I_\varphi(\mathbb{P})$, obtained by the above "duality" technique, equals the direct plug-in one*

$$\widehat{I}_\varphi^{emp} := I_\varphi(\widehat{\mathbb{P}}) = \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \varphi\left(\frac{\widehat{p}_{x,y}}{\widehat{p}_x\widehat{p}_y}\right)\widehat{p}_x\widehat{p}_y. \tag{26}$$

Indeed, we have by its proper definition

$$\widehat{I}_\varphi = \sup_{\theta\in\Theta} M_n(\theta), \text{ where } M_n(\theta) = \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \left[\varphi'(e^{\theta_{x,y}})\widehat{p}_{x,y} - e^{\theta_{x,y}}\varphi'(e^{\theta_{x,y}})\widehat{p}_x\widehat{p}_y + \varphi(e^{\theta_{x,y}})\widehat{p}_x\widehat{p}_y\right]. \tag{27}$$

Differentiating (27) with respect to $\theta_{x,y}$ for $(x,y)\in\mathcal{X}\times\mathcal{Y}$ yields

$$\frac{\partial}{\partial\theta_{x,y}}M_n(\theta) = \varphi''(e^{\theta_{x,y}})\left(e^{\theta_{x,y}}\,\widehat{p}_{x,y} - e^{2\theta_{x,y}}\,\widehat{p}_x\widehat{p}_y\right).$$

Canceling derivatives $\frac{\partial}{\partial\theta_{x,y}}M_n(\theta)$ yields

$$\widehat{\theta}_{x,y} = \log\frac{\widehat{p}_{x,y}}{\widehat{p}_x\widehat{p}_y}, \quad (x,y)\in\mathcal{X}\times\mathcal{Y},$$

which is independent from the choice of φ for this particular model. Finally, straightforward simplifications yield

$$\widehat{I}_\varphi = M_n(\widehat{\theta}) = \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \varphi\left(\frac{\widehat{p}_{x,y}}{\widehat{p}_x\widehat{p}_y}\right)\widehat{p}_x\widehat{p}_y = \widehat{I}_\varphi^{emp}.$$

Particularly, for $\varphi(x) = \varphi_2(x) := (x-1)^2/2$, the estimate $\widehat{I}_{\varphi_2}$ of the χ^2-mutual information – or χ^2 measure of independence – obtained by the duality technique is shown to equal (up to the factor $2n$) the classical χ^2 statistics. Hence, in the context of finite-discrete distributions, using the exponential model described in Example 2.3, we see that the proposed approach, via duality technique, recovers the classical direct plug-in one, in particular, the well-known classical χ^2-independence test.

Remark 2.7. *For finite discrete distributions (with known support, of size say K, see Example 2.3), as in plug-in estimation of Shannon entropy (see e.g. Chao and Shen (2003)), the direct plug-in estimates $\widehat{I}_\varphi^{emp}$ are valid with small bias if the sample size $n >> K$. If the sample size n is not sufficiently large compared to the space size K, models $h_\theta(\cdot)$ other than (12) should be used (through e.g. the model selection procedure described in Section 2.4), with small parameter dimension, and the corresponding dual estimate $\widehat{I}_\varphi$, if the model $h_\theta(\cdot)$ is correctly specified, could be more promising than the direct plug-in one $\widehat{I}_\varphi^{emp}$.*

Example 2.8. *Note that when dealing with semiparametric copula models*

$$h_\theta(x,y) = c_\theta(F_1(x), F_2(y)),$$

with unknown nonparametric cumulative distribution functions F_1 and F_2, it is necessary to estimate them, using for example their empirical counterparts. Denote by $\widehat{F}_1(\cdot)$ and $\widehat{F}_2(\cdot)$ the empirical cumulative distribution functions associated, respectively, to the samples $X_1 \ldots, X_n$ and $Y_1, \ldots, Y_n$, i.e.,

$$\widehat{F}_1(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{]-\infty,x]}(X_i) \quad \text{and} \quad \widehat{F}_2(y) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{]-\infty,y]}(Y_i).$$

So that $\widehat{I}_\varphi$ and $\widehat{\theta}_\varphi$ become

$$\widehat{I}_\varphi = \sup_{\theta\in\Theta} \left\{ \frac{1}{n}\sum_{i=1}^{n} \varphi'\left(c_\theta\left(\widehat{F}_1(X_i), \widehat{F}_2(Y_i)\right)\right) - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \varphi^*\left(\varphi'\left(c_\theta\left(\widehat{F}_1(X_i), \widehat{F}_2(Y_j)\right)\right)\right) \right\}$$

$$\widehat{\theta}_\varphi = \arg\sup_{\theta\in\Theta} \left\{ \frac{1}{n}\sum_{i=1}^{n} \varphi'\left(c_\theta\left(\widehat{F}_1(X_i), \widehat{F}_2(Y_i)\right)\right) - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \varphi^*\left(\varphi'\left(c_\theta\left(\widehat{F}_1(X_i), \widehat{F}_2(Y_j)\right)\right)\right) \right\}.$$

Note that $n\widehat{F}_1(X_i)$ is the rank of X_i in the sample $X_1, \ldots, X_n$ and $n\widehat{F}_2(X_j)$ is the rank of Y_j in the sample $Y_1, \ldots, Y_n$. For some copula models, the copula density $c_\theta(u_1, u_2)$ may be unbounded when either u_1 or u_2 tends to 1; see e.g. Genest et al. (1995). In this case, to avoid this difficulty, the "rescaled" empirical cumulative distribution functions

$$\widetilde{F}_1(\cdot) := \frac{n}{n+1}\widehat{F}_1(\cdot), \quad \widetilde{F}_2(\cdot) := \frac{n}{n+1}\widehat{F}_2(\cdot)$$

should be used instead of the standard ones $\widehat{F}_1(\cdot)$ and $\widehat{F}_2(\cdot)$.

2.4. **A model selection procedure for the ratio $\mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{P}^\perp$ through φ-MI criterion.** Let $\mathcal{M}_{\Theta_1} := \left\{h_{\theta_1,1}(\cdot,\cdot);\, \theta_1 \in \Theta_1 \subset \mathbb{R}^{d_1}\right\}, \ldots, \mathcal{M}_{\Theta_L} := \left\{h_{\theta_L,L}(\cdot,\cdot);\, \theta_L \in \Theta_L \subset \mathbb{R}^{d_L}\right\}$ be L candidate models for the ratio $\mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{P}^\perp$. For any model $\mathcal{M}_{\Theta_\ell}$, denote by $\widehat{\theta}_\ell$ the estimate of θ_T given by

$$\widehat{\theta}_\ell := \arg\sup_{\theta_\ell\in\Theta_\ell} M_n(\theta_\ell).$$

The corresponding "expected" criterion is

$$M(\widehat{\theta}_\ell) = \int_{\mathcal{X}\times\mathcal{Y}} f_{\widehat{\theta}_\ell}(x,y)\,\mathrm{d}\mathbb{P}(x,y) - \int_{\mathcal{X}\times\mathcal{Y}} g_{\widehat{\theta}_\ell}(x,y)\,\mathrm{d}\mathbb{P}(x,y)^\perp.$$

From the representation (23), we can see that the larger the expected criterion $M(\widehat{\theta}_\ell)$ of the model is, the closer the model is to the true one. We propose then the following k-fold cross-validation procedure for model selection using the proposed estimate (24) of φ-MI.

(1) Partition the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ into k equal size (n_k) subsamples. (Denote the i-th subsample $(X_{(i-1)n_k+1}, Y_{(i-1)n_k+1}), \ldots, (X_{in_k}, Y_{in_k})$, for all $i = 1, \ldots, k$);
(2) Consider a candidate model $\mathcal{M}_{\Theta_\ell}$;
(3) From the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ remove the i-th subsample; compute the estimate $\widehat{\theta}_\ell^{(-i)}$ given by (25) using the remaining $n - n_k$ observations, i.e.,

$$\widehat{\theta}_\ell^{(-i)} = \arg \sup_{\theta_\ell \in \Theta_\ell} M_{n-n_k}(\theta_\ell);$$

(4) Repeat steps (2) and (3) for all $i = 1, \ldots, k$, and obtain the following "estimate"

$$C_V(\mathcal{M}_{\Theta_\ell}) := \frac{1}{k} \sum_{i=1}^{k} \left(\frac{1}{n_k} \sum_{j=(i-1)n_k+1}^{in_k} f_{\widehat{\theta}_\ell^{(-i)}}(X_j, Y_j) - \frac{1}{n_k^2} \sum_{j,m=(i-1)n_k+1}^{in_k} g_{\widehat{\theta}_\ell^{(-i)}}(X_j, Y_m) \right)$$

of the expected criterion $M(\widehat{\theta}_\ell)$, i.e.,

$$M(\widehat{\theta}_\ell) = \int_{\mathcal{X}\times\mathcal{Y}} f_{\widehat{\theta}_\ell}(x, y) \, \mathrm{d}\mathbb{P}(x, y) - \int_{\mathcal{X}\times\mathcal{Y}} g_{\widehat{\theta}_\ell}(x, y) \, \mathrm{d}\mathbb{P}(x, y)^{\perp};$$

(5) Repeat steps (2-4) for all $\ell = 1, \ldots, L$, and select the "optimal" model $\mathcal{M}_{\Theta_{\ell^*}}$ that maximizes $C_V(\mathcal{M}_{\Theta_\ell})$ over $\ell = 1, \ldots, L$, i.e., the model $\mathcal{M}_{\Theta_{\ell^*}}$ with

$$\ell^* := \arg \sup_{\ell \in \{1,\ldots,L\}} C_V(\mathcal{M}_{\Theta_\ell}).$$

Other model selection-type procedures can be investigated, through e.g. correcting the bias of $M_n(\widehat{\theta}_\ell)$ as an estimate of the expected criterion $M(\widehat{\theta}_\ell)$, and selecting the model that maximizes the obtained information criterion corrected from bias. The correction can be made e.g. by asymptotic evaluation of the bias as in classical AIC criterion, or using bootstrap; see e.g. Konishi and Kitagawa (2008) and Shao and Tu (1995).

3. Asymptotic properties of the estimates

We state in Section 3.1 the consistency of both estimates $\widehat{I}_\varphi$ and $\widehat{\theta}_\varphi$, of the φ-MI and the parameter θ_T. Section 3.2 gives, under the null hypothesis of independence, the limiting distribution of the estimate $\widehat{I}_{\varphi_1}$ of the KL-MI, as well as the corresponding estimate $\widehat{\theta}_{\varphi_1}$ of the parameter θ_T, for some specific forms of the model $\{h_\theta(\cdot, \cdot);\, \theta \in \Theta\}$. Section 3.3 provides bootstrap calibration of the critical value of any $\widehat{I}_\varphi$-based test statistic for general forms of the model $\{h_\theta(\cdot, \cdot);\, \theta \in \Theta\}$.

3.1. **Consistency.** In this section, we state consistency of the estimate $\widehat{I}_\varphi$, of the φ-MI, defined by (17), as well as the consistency of the estimates $\widehat{\theta}_\varphi$ of θ_T. We will use classical techniques from M-estimation theory. We will make use of the following conditions.

(A.5) The parameter space Θ is a compact subset of $\mathbb{R} \times \mathbb{R}^d$;
(A.6) $\int_{\mathcal{X}\times\mathcal{Y}} \sup_{\theta\in\Theta} |f_\theta(x,y)| \, \mathrm{d}\mathbb{P}(x,y) < \infty$;
(A.7) $\int_{\mathcal{X}\times\mathcal{Y}} \sup_{\theta\in\Theta} g_\theta(x,y)^2 \, \mathrm{d}\mathbb{P}^\perp(x,y) < \infty$,

where f_θ and g_θ are defined respectively by (19) and (20). Note that assumptions (A.6-7) imply (A.3-4).

Proposition 3.1. *Assume that conditions (A.1, 5-7) hold. Then, the estimates $\widehat{I}_\varphi$ of $I_\varphi(\mathbb{P})$ defined by (17) and the estimates $\widehat{\theta}_\varphi$ of θ_T defined by (18) are consistent. Precisely, as $n \to \infty$, the following convergences in probability hold*

$$\widehat{I}_\varphi \to I_\varphi(\mathbb{P}) \quad \textit{and} \quad \widehat{\theta}_\varphi \to \theta_T.$$

Remark 3.2. *Since in practice, all models are generally "misspecified", the true parameter value θ_T may not exist, it can however be replaced by the "pseudo-true" value $\theta_T^* := \arg\sup_{\theta\in\Theta} M(\theta)$, and the results of consistency in the above proposition remain valid.*

3.2. **The limiting distribution of the estimate $\widehat{I}_{\varphi_1}$ of KL-MI.** We will give now the limiting distribution of the particular statistical test based on the estimate $\widehat{I}_{\varphi_1}$ of classical KL-MI, for specific forms of the model $h_\theta(\cdot,\cdot)$, under the null hypothesis of independence $\mathcal{H}_0 : \mathbb{P} = \mathbb{P}^\perp$. Consider the following specific form of the model $h_\theta(\cdot,\cdot)$

$$h_\theta(x,y) = \exp\left(\alpha + m_\beta(x,y)\right) \quad \text{with} \quad m_\beta(x,y) := \sum_{k=1}^{d} \beta_k \xi_k(x)\zeta_k(y), \tag{28}$$

for some specified measurable real valued functions ξ_k and ζ_k, $k = 1,\ldots,d$, defined, respectively, on $\mathcal{X}$ and $\mathcal{Y}$. The parameter θ is the vector $\theta := (\alpha, \beta_1, \ldots, \beta_d)^\top \in \Theta \subset \mathbb{R} \times \mathbb{R}^d$. In this case, the functions (19) and (20) become

$$f_\theta(x,y) = \alpha + \sum_{k=1}^{d} \beta_k \xi_k(x)\zeta_k(y)$$

and

$$g_\theta(x,y) = \exp\left(\alpha + \sum_{k=1}^{d} \beta_k \xi_k(x)\zeta_k(y)\right) - 1.$$

The value θ_0, corresponding to the independence, here is $\theta_0 = \mathbf{0} := (0,\ldots,0)^\top \in \mathbb{R}^{1+d}$. We will give the limiting distributions of $\widehat{\theta}_{\varphi_1}$ and $\widehat{I}_{\varphi_1}$, under the null hypothesis of independence $\mathbb{P} = \mathbb{P}^\perp$, i.e., when $\theta_T = \theta_0 = \mathbf{0}$. We will consider the following assumptions.

(A.8) There exists a neighborhood $N(\theta_T)$ of θ_T such that the third order partial derivative functions $\{(x,y) \mapsto (\partial^3/\partial^3\theta) f_\theta(x,y);\ \theta \in N(\theta_T)\}$
(resp. $\{(x,y) \mapsto (\partial^3/\partial^3\theta) g_\theta(x,y);\ \theta \in N(\theta_T)\}$) are dominated by some functions $\mathbb{P}$-integrable (resp. some function $\mathbb{P}^\perp$-square-integrable);

(A.9) The integrals $\mathbb{P}\left\|f'_{\theta_T}\right\|^2$, $\mathbb{P}^\perp\left\|g'_{\theta_T}\right\|^2$, $\mathbb{P}\left\|f''_{\theta_T}\right\|$, $\mathbb{P}^\perp\left\|g''_{\theta_T}\right\|^2$ exist, and the matrix

$$\Sigma_1 := -\left(\mathbb{P} f''_{\theta_T} - \mathbb{P}^\perp g''_{\theta_T}\right) \tag{29}$$

is nonsingular.

Theorem 3.3. *Assume that conditions (A.1-2,5-9) hold and that $\mathbb{P} = \mathbb{P}^\perp$ (i.e., $\theta_T = \mathbf{0}$). Then,*

(a) $\sqrt{n}\,\widehat{\theta}_{\varphi_1}$ converges in distribution to a centered multivariate normal random variable with covariance matrix $\Sigma = \Sigma_1^{-1}\Sigma_2\Sigma_1^{-1}$, where Σ_1 and Σ_2 are given respectively by (29) and (40);

(b) $2n\,\widehat{I}_{\varphi_1}$ converges in distribution to the random variable $Z^\top Z$, where Z is a centered multivariate normal random variable with covariance matrix

$$C = \Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}.$$

Remark 3.4. *For the finite-discrete case, using the modeling (13) in Example 2.3, we can see that the corresponding matrix Σ_2 is of rank $(K_1-1)(K_2-1)$ and that the limiting distribution of $2n\widehat{I}_\varphi = 2n\widehat{I}^{emp}_\varphi$ is a χ^2-distribution with $(K_1-1)(K_2-1)$ degrees of freedom, in particular, we recover the classical χ^2-independence test theorem (for the case of finite-discrete distributions).*

3.3. **Bootstrap calibration.** In the general context of model (9), for a given φ-MI, we propose the following bootstrap procedure to calibrate the critical value of the corresponding test statistic. The critical value, denote it b_α, is the upper α-quantile of the distribution of the test statistic $S_n := 2n\widehat{I}_\varphi$, under the null hypothesis $\mathcal{H}_0$ of independence.

(1) Generate bootstrap sample $(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$ from the product empirical distribution $\widehat{\mathbb{P}}^\perp = \widehat{\mathbb{P}}_1 \otimes \widehat{\mathbb{P}}_2$ of the original sample $(X_1, Y_1), \ldots, (X_n, Y_n)$;
(2) Compute the value of the statistic $S_n^* := 2n\widehat{I}_\varphi^*$ from the bootstrap sample;
(3) Repeat steps (1) and (2) $B = 1000$ times, independently, to obtain the realizations $\left\{S_{n,1}^*, S_{n,2}^*, \ldots, S_{n,B}^*\right\}$;
(4) Estimate b_α by $\widetilde{b}_\alpha :=$ the $(1-\alpha)$th quantile of the sequence $\left\{S_{n,1}^*, S_{n,2}^*, \ldots, S_{n,B}^*\right\}$.

4. Large deviations principle and Bahadur asymptotic efficiency

In this section, we compare Bahadur asymptotic efficiency of φ-MI based independence tests and show that the test based on classical Kullback-Leibler mutual information is the most efficient. Given $(\widehat{I}_{\varphi_1})_n$ and $(\widehat{I}_{\varphi_2})_n$ two sequences of statistics, for the test problem (5), numbers

$\alpha \in (0,1)$, $\gamma \in (0,1)$ and an alternative hypothesis $\mathbb{P} \neq \mathbb{P}^{\perp}$, we define $n_i(\alpha,\gamma,\mathbb{P})$, for $i \in \{1,2\}$, respectively, as the minimal number of observations needed for the test based on $\widehat{I}_{\varphi_i}$ to have signification level α and power level γ. Then, Bahadur asymptotic relative efficiency of $(\widehat{I}_{\varphi_1})_n$ with respect to $(\widehat{I}_{\varphi_2})_n$ is defined as (if the limit exists)

$$\lim_{\alpha \to 0} \frac{n_2(\alpha,\gamma,\mathbb{P})}{n_1(\alpha,\gamma,\mathbb{P})}.$$

It is well known, see for example Nikitin (1995) and van der Vaart (1998) Chapter 14, that if both sequences $(\widehat{I}_{\varphi_1})_n$ and $(\widehat{I}_{\varphi_2})_n$ satisfy a large deviation principle under the null hypothesis (with good rate functions $e_{\varphi_1}(\cdot)$ and $e_{\varphi_2}(\cdot)$) and also a law of large number under a given alternative hypothesis $\mathcal{H}_1 : \mathbb{P} \neq \mathbb{P}^{\perp}$, with asymptotic means $\mu_{\varphi_1}(\mathbb{P})$ and $\mu_{\varphi_2}(\mathbb{P})$, respectively, then the Bahadur asymptotic relative efficiency equals $e_{\varphi_1}(\mu_{\varphi_1}(\mathbb{P}))/e_{\varphi_2}(\mu_{\varphi_2}(\mathbb{P}))$. Particularly, the most efficient test maximizes Bahadur slope $e_{\varphi}(\mu_{\varphi}(\mathbb{P}))$. A law of large number under the alternative hypothesis is given for the sequence $(\widehat{I}_{\varphi})_n$ in Proposition 3.1 above; the excepted value $\mu_{\varphi}(\mathbb{P})$ being $\mu_{\varphi}(\mathbb{P}) = I_{\varphi}(\mathbb{P}) = D_{\varphi}(\mathbb{P},\mathbb{P}^{\perp})$. The following theorem establishes a large deviation principle under the null hypothesis of independence. It relies on some generalization due to Eichelsbacher and Schmock (2002) of classical Sanov theorem to finer topologies and the contraction principle. Let $\mathcal{G}$ be the set of measurable functions, from $\mathcal{X} \times \mathcal{Y}$ into $\mathbb{R}$, given by

$$\mathcal{G} := \mathcal{B} \cup \{\varphi'(h_\theta); \theta \in \Theta\} \cup \{\varphi^*(\varphi'(h_\theta)); \theta \in \Theta\},$$

where $\mathcal{B}$ is the set of all measurable bounded functions from $\mathcal{X} \times \mathcal{Y}$ into $\mathbb{R}$. Recall that $\mathcal{M}_1 = \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ is the set of all probability measures on $\mathcal{X} \times \mathcal{Y}$, and let us introduce the subset

$$\mathcal{M}_{\mathcal{G}} := \mathcal{M}_{\mathcal{G}}(\mathcal{X}\times\mathcal{Y}) := \left\{ P \in \mathcal{M}_1 : \int_{\mathcal{X}\times\mathcal{Y}} |\varphi'(h_\theta)|\, dP < \infty, \int_{\mathcal{X}\times\mathcal{Y}} |\varphi^*(\varphi'(h_\theta))|\, dP^{\perp} < \infty, \forall \theta \in \Theta \right\}.$$

Define on $\mathcal{M}_{\mathcal{G}}$ the $\tau_{\mathcal{G}}$-topology as the coarsest one that makes applications $P \in \mathcal{M}_{\mathcal{G}} \mapsto \int_{\mathcal{X}\times\mathcal{Y}} \varphi'(h_\theta)\, dP$, $P \in \mathcal{M}_{\mathcal{G}} \mapsto \int_{\mathcal{X}\times\mathcal{Y}} f\, dP$, $P \in \mathcal{M}_{\mathcal{G}} \mapsto \int_{\mathcal{X}\times\mathcal{Y}} \varphi^*(\varphi'(h_\theta))\, dP^{\perp}$ and $P \in \mathcal{M}_{\mathcal{G}} \mapsto \int_{\mathcal{X}\times\mathcal{Y}} f\, dP^{\perp}$ continuous, for all $\theta \in \Theta$ and all $f \in \mathcal{B}$. Finally, define, for all $Q \in \mathcal{M}_{\mathcal{G}}$, the "pseudo-divergence"

$$\mathcal{D}_{\varphi}(Q,Q^{\perp}) := \sup_{\theta\in\Theta} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} \varphi'(h_\theta(x,y))\, dQ(x,y) - \int_{\mathcal{X}\times\mathcal{Y}} \varphi^*\left(\varphi'(h_\theta(x,y))\right) dQ^{\perp}(x,y) \right\}.$$

Obviously, $\mathcal{D}_{\varphi}(Q,Q^{\perp}) \leq D_{\varphi}(Q,Q^{\perp}) =: I_{\varphi}(Q)$ with equality for probability distributions such that $dQ/dQ^{\perp} = h_\theta$ for some $\theta \in \Theta$. Note also that $Q \in \mathcal{M}_{\mathcal{G}} \mapsto \mathcal{D}_{\varphi}(Q,Q^{\perp})$ is continuous with respect to the $\tau_{\mathcal{G}}$-topology as the supremum over the compact set Θ of continuous functions.

The large deviation principle for the sequence $(\widehat{\mathbb{P}}(\cdot))_n$ of empirical measures defined by (8), established by Eichelsbacher and Schmock (2002), requires the existence of exponential moments; in the context of the model (9), we thus assume

(A.10) for all $f \in \mathcal{G}$, for all $a > 0$,

$$\int_{\mathcal{X}\times\mathcal{Y}} \exp(a|f|)\, \mathrm{d}\mathbb{P} < \infty.$$

Note that the strong assumption $(A.10)$ implies (A.3-4) if $\mathbb{P} = \mathbb{P}^{\perp}$. In the context of the models described in Examples 2.1 to 2.5, assumption (A.10) may not be satisfied for some φ-divergences ; particularly, it does not generally hold for power-divergences (except for finite-discrete distribution models described in Example 2.3). A sufficient condition for (A.10) is

(A.11) there exist real numbers $m, M \in (a_\varphi, b_\varphi)$ such that $m < h_\theta(x,y) < M, \quad \forall (x,y) \in \mathcal{X}\times\mathcal{Y}, \forall \theta \in \Theta$.

Indeed, for all $a > 0$, the functions $\exp(a|\varphi'(h_\theta)|)$ and $\exp(a|\varphi^*(\varphi'(h_\theta))|)$ are bounded and therefore integrable with respect to both $\mathbb{P}$ and $\mathbb{P}^{\perp}$. Again, (A.11) is not generally satisfied for models described in the previous examples for power-divergences, but it may be artificially verified by truncating the distributions in the models. Let us also point out that Theorems 4.1 and 4.2 below may remain true with some alternative assumptions on the distribution queues, lighter than (A.10). Particularly, simulations performed in Section 5 for bivariate Gaussian distributions tend to show that Theorem 4.2 holds for the Gaussian model described in Example 2.1. For getting a closed form for the LDP of $(\widehat{I}_\varphi)_n$, we will establish the right-continuity of the rate function, making use of one of the following assumptions:

(A.12.a) (X,Y) is finite-discrete, supported by $\mathcal{X}\times\mathcal{Y}$;

(A.12.b) The model $\left\{h_\theta(\cdot,\cdot);\, \theta = (\alpha, \beta^\top)^\top \in \Theta\right\}$ is of the from $h_\theta(x,y) = \exp\left(\alpha + m_\beta(x,y)\right)$ with the condition that, for any constant c and any β, we have $\mathbb{P}^{\perp}\left(m_\beta(X,Y) = c\right) \neq 0$ iff $\beta = (0,\ldots,0)^\top$ and $c = 0$.

Theorem 4.1. *Let (X,Y) be a couple of independent random variables with joint distribution $\mathbb{P} = \mathbb{P}^{\perp} \in \mathcal{M}_\Theta \cap \mathcal{M}_\mathcal{G}$.*

(1) Suppose that conditions (A.1-2, 5-7, 10 and 12.b) are satisfied. Then, the sequence $(\widehat{I}_\varphi)_n$ of estimates, of $I_\varphi(\mathbb{P}) = 0$, given by (17), satisfies the following large deviation principle

$$\frac{1}{n}\log \mathbb{P}^{\perp}\left(\widehat{I}_\varphi > d\right) \overset{n\to\infty}{\longrightarrow} -e_\varphi(d), \quad d > 0, \tag{30}$$

where the good rate function $e_\varphi(\cdot)$ is

$$e_\varphi(d) := \inf_{Q\in\Omega_d} \mathbb{K}(Q, \mathbb{P}^{\perp}) \;\; \textit{with} \;\; \Omega_d := \left\{Q \in \mathcal{M}_\mathcal{G} \;\textit{such that}\; \mathcal{D}_\varphi(Q, Q^{\perp}) \geq d\right\}. \tag{31}$$

(2) Assume that conditions (A.1-2, 5 and 12.a) are satisfied. Then the above statement holds if $\mathcal{M}_{\mathcal{G}}$ is replaced by the set of all discrete-finite distributions with the same finite support $\mathcal{X} \times \mathcal{Y}$.

In view of Proposition 3.1 and Theorem 4.1 above, the Bahadur slope of the independence test based on $\widehat{I}_{\varphi}$, for any φ, is given then by

$$\begin{aligned} s_{\varphi} &:= e_{\varphi}(I_{\varphi}(\mathbb{P})) \\ &= \inf\{\mathbb{K}(Q, \mathbb{P}^{\perp}) : \mathcal{D}_{\varphi}(Q, Q^{\perp}) \geq D_{\varphi}(\mathbb{P}, \mathbb{P}^{\perp})\}. \end{aligned}$$

Since $\mathcal{D}_{\varphi}(\mathbb{P}, \mathbb{P}^{\perp}) = D_{\varphi}(\mathbb{P}, \mathbb{P}^{\perp})$, we have $\mathbb{P} \in \{Q : \mathcal{D}_{\varphi}(Q, Q^{\perp}) \geq D_{\varphi}(\mathbb{P}, \mathbb{P}^{\perp})\}$, so that, for any φ,

$$s_{\varphi} \leq \mathbb{K}(\mathbb{P}, \mathbb{P}^{\perp}) = I_{KL}(\mathbb{P}) = I_{\varphi_1}(\mathbb{P}). \tag{32}$$

Equality is achieved in (32) for the divergence $D_{\varphi} = \mathbb{K}$. Indeed,

$$s_{KL} = \inf\{\mathbb{K}(Q, \mathbb{P}^{\perp}) : \mathcal{D}_{KL}(Q, Q^{\perp}) \geq \mathbb{K}(\mathbb{P}, \mathbb{P}^{\perp})\}.$$

Straightforward computations yield

$$\mathbb{K}(Q, \mathbb{P}^{\perp}) = \mathbb{K}(Q, Q^{\perp}) + \mathbb{K}(Q_1, \mathbb{P}_1) + \mathbb{K}(Q_2, \mathbb{P}_2),$$

for any $Q \in \mathcal{M}_{\mathcal{G}}$. Particularly, for any Q such that $\mathcal{D}_{KL}(Q, Q^{\perp}) \geq \mathbb{K}(\mathbb{P}, \mathbb{P}^{\perp})$, we have $\mathbb{K}(Q, Q^{\perp}) \geq \mathcal{D}_{KL}(Q, Q^{\perp}) \geq \mathbb{K}(\mathbb{P}, \mathbb{P}^{\perp})$, hence,

$$\mathbb{K}(Q, \mathbb{P}^{\perp}) \geq \mathbb{K}(Q, Q^{\perp}) \geq \mathbb{K}(\mathbb{P}, \mathbb{P}^{\perp}),$$

so that

$$s_{KL} \geq \mathbb{K}(\mathbb{P}, \mathbb{P}^{\perp}). \tag{33}$$

Combining (32) and (33), we obtain

Theorem 4.2. *Let (X, Y) be a couple of random variables with joint distribution $\mathbb{P} \in \mathcal{M}_{\Theta} \cap \mathcal{M}_{\mathcal{G}}$. Suppose that either conditions (A.1-2, 5-7, 10 and 12.b) or (A.1-2, 5 and 12.a) are satisfied. For the test problem (5), the test based on the estimate $\widehat{I}_{\varphi_1}$, see (17), of the Kullback-Leibler mutual information, is uniformly (i.e., whatever be the alternative $\mathbb{P} \neq \mathbb{P}^{\perp}$) the most efficient test, in Bahadur sense, among all $\widehat{I}_{\varphi}$-based tests, including the classical χ^2-independence one.*

Remark 4.3. *Assume that $\mathbb{P}$ is a finite-discrete distribution. We obtain then that KL-MI based independence test is more efficient than the classical χ^2 independence one. This result was already stated, in goodness-of-fit testing for finite-discrete distributions, see e.g. van der Vaart (1998) Chapter 17 Section 17.6. The above theorem extends it to testing independence, for more general probability distributions, not necessarily finite-discrete.*

5. Simulations

This Section aims at numerically comparing through simulations φ-MI based tests with other independence or non-correlation tests. Precisely, Section 5.1 focuses on finite-discrete random vectors, for which the optimal KL-MI test is compared to the very popular (but not optimal) χ^2-independence test. Section 5.2 compares KL-MI and χ^2 tests to classical non-correlation tests of Pearson, Kendall and Spearman. Finally, Section 5.3 deals with the example of the copula density model of Farlie-Gumbel-Morgenstern (FGM), for which the critical values of KL-MI and χ^2-MI tests are derived through the bootstrap procedure described in Section 3.3.

5.1. **Testing independence of finite-discrete random variables.** As stated in Example 2.6, the dual estimates $\widehat{I}_\varphi$ given by (17) equal the direct empirical ones (26). Their properties and asymptotic behavior are well-known; see e.g. Pardo (2006). They are recovered by Propositions 3.1, Theorem 3.3 and Theorem 4.2. We illustrate these properties through simulations, by comparing the power of KL-MI and χ^2-MI tests, for various sample sizes and finite-discrete supports $\mathcal{X} = \mathcal{Y} = \{1, \ldots, K\}$, and for alternatives $P \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ of the form $P_\theta := (p_{x,y;\theta})_{(x,y)}$, with

$$p_{x,y;\theta} = (1-\theta)\frac{1}{K^2} + \theta\frac{1}{K}\mathbb{1}_{\{x=y\}}, \quad (x,y) \in \mathcal{X} \times \mathcal{Y}, \tag{34}$$

where $K = |\mathcal{X}| = |\mathcal{Y}|$ and $\theta \in (0,1)$, i.e., the random variables X and Y are uniformly distributed on the set $\{1, \ldots, K\}$, and the conditional distribution $P_{Y|X=x}(\cdot)$, of Y knowing $X = x$, is the mixture of the uniform distribution on $\{1, \ldots, K\}$ with weight $(1-\theta)$ and the Dirac measure $\delta_x(\cdot)$ with weight θ, for all $x \in \{1, \ldots, K\}$. Hence, for $\theta = \theta_0 = 0$, X and Y are independent, while for $\theta = 1$, we have $Y = X$. The level of the tests has been set to $\alpha = 0.01$. The asymptotic distribution of $2n\widehat{I}_\varphi$ is $\chi^2\left((K-1)(K-1)\right)$, a χ^2-distribution with $(K-1)^2$ degrees of freedom, for both KL-MI or χ^2-MI. The critical value $b_{0.01}$ of both test statistics is taken then to be the upper 0.01-quantile of the $\chi^2\left((K-1)(K-1)\right)$-distribution. Then, we have estimated their respective powers, by means of Monte-Carlo procedure from 10000 samples drawn according to P_θ given by (34), for various mixture parameter values $\theta \in (0,1)$. The results are presented in Table 2, Figure 1 and Figure 2. We can see that the KL-MI test outperforms the classical χ^2 one. The nominal levels of both KL-MI and χ^2-MI test statistics are both close to the test level $\alpha = 0.01$.

5.2. **Comparison of φ-MI based and noncorrelation tests in the Gaussian setting.** For bidimensional normally distributed random vectors, the corresponding model $h_\theta(\cdot,\cdot)$, see Example 2.1, is of the form (28), so that the asymptotic distribution of the dual KL-MI based test statistic $2n\widehat{I}_{\varphi_1}$ is explicit. Hence, explicit (asymptotic) critical value can be obtained for

$K = \lvert\mathcal{X}\rvert = \lvert\mathcal{Y}\rvert = 2$	$\theta =$	0	0.08	0.18	0.28	0.38	0.48	0.58	0.68
$n = 30$	KL-MI test power	0.0123	0.0242	0.0647	0.1681	0.3343	0.5690	0.7981	0.9415
	χ^2 test power	0.0102	0.0200	0.0550	0.1433	0.2968	0.5330	0.7703	0.9288
$n = 40$	KL-MI test power	0.0119	0.0213	0.0764	0.2176	0.4502	0.7180	0.9046	0.9850
	χ^2 test power	0.0100	0.0184	0.0694	0.2006	0.4272	0.6970	0.8957	0.9839
$K = \lvert\mathcal{X}\rvert = \lvert\mathcal{Y}\rvert = 3$	$\theta =$	0	0.07	0.15	0.23	0.31	0.39	0.47	0.55
$n = 35$	KL-MI test power	0.0192	0.0261	0.0604	0.1503	0.3162	0.5267	0.7476	0.8952
	χ^2 test power	0.0081	0.0118	0.0371	0.1157	0.2708	0.4878	0.7259	0.8895
$n = 50$	KL-MI test power	0.0152	0.0261	0.0782	0.2152	0.4369	0.7150	0.9039	0.9816
	χ^2 test power	0.0088	0.0167	0.0648	0.1929	0.4283	0.7124	0.9057	0.9832

TABLE 2. Comparison of powers of KL-MI and χ^2-MI tests. The number of cells K is indicated at the top left of each block. The sample sizes n are given by the first column while the mixture parameter values θ, see its definition in (34), are given by the first row.

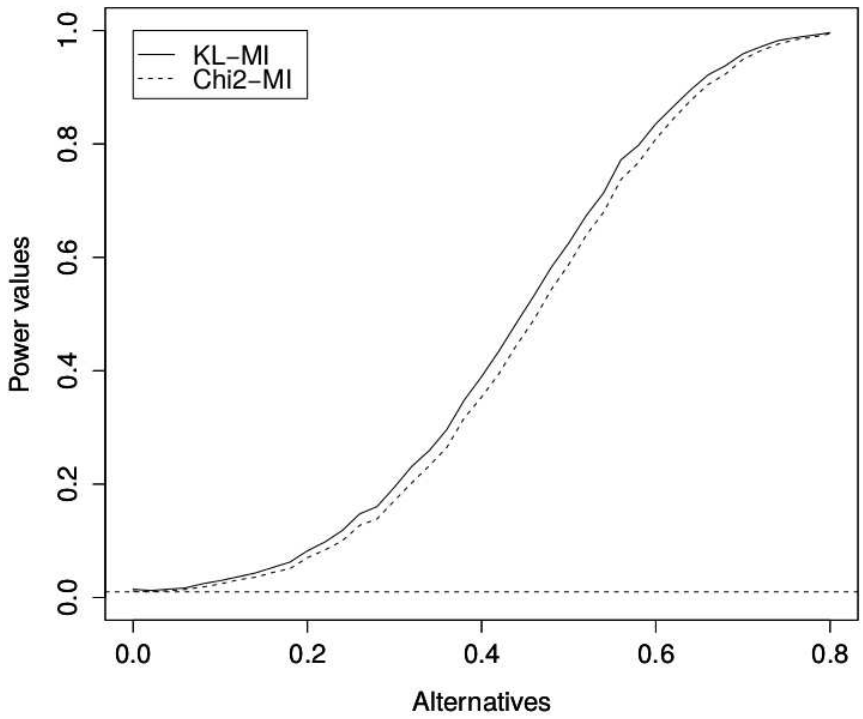

FIGURE 1. Comparison of KL-MI and χ^2-MI based tests for finite-discrete random variables taking values in $\{1, 2\}$, with $n = 30$.

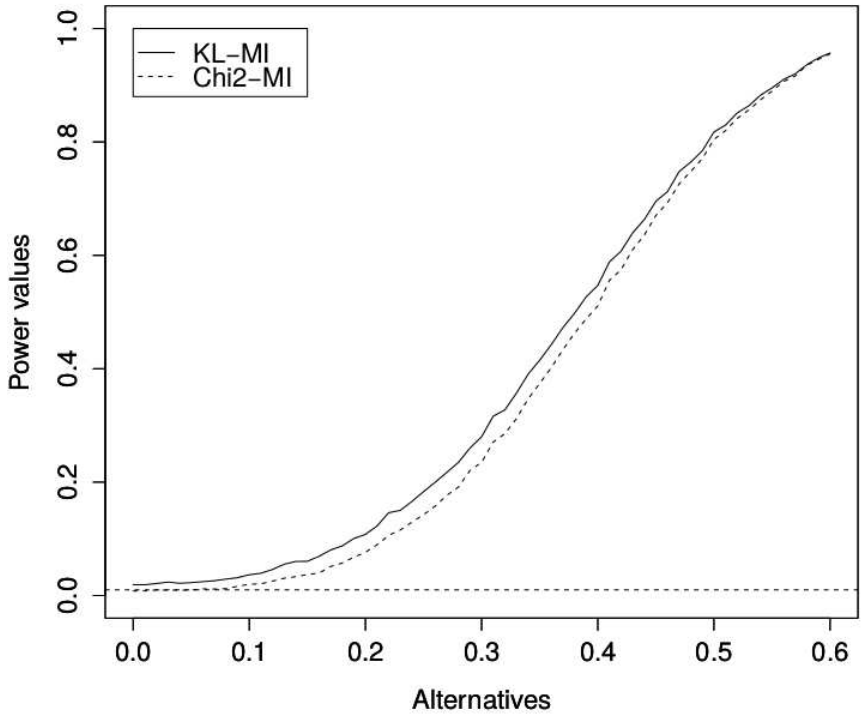

FIGURE 2. Comparison of KL-MI and χ^2-MI based tests for finite-discrete random variables taking values in $\{1, 2, 3\}$, with $n = 35$.

the test statistic $2n\widehat{I}_{\varphi_1}$. Although assumption (A.10) may not be satisfied without restricting the support of (X, Y) to a bounded subset of $\mathbb{R}^2$, we can compare numerically the powers of the φ-MI based tests. Precisely, in this Section we manage to compare the powers of KL-MI and χ^2-MI independence tests with noncorrelation tests for samples of size $n = 50$ drawn according to bivariate normal distributions. We have fixed the level $\alpha = 0.05$ and computed the critical value of KL-MI based test by means of Monte-Carlo simulations of the asymptotic distribution of $2n\widehat{I}_{\varphi_1}$ given by Theorem 3.3 (10000 samples of the variable Z in Theorem 3.3 have been simulated; the critical value has been obtained as the 0.95-quantile of the linearly interpolated empirical cumulative density function). The critical value for the χ^2-MI based test have been estimated directly by simulating 10000 samples of size 50 of a bivariate Gaussian random vector with independent centered and reduced distribution and computing the 0.95-quantile of the corresponding tes statistic $2n\widehat{I}_{\varphi_2}$. Then we have estimated the power of these tests as well as noncorrelation tests of Pearson, Spearman and Kendall, still by Monte-Carlo methods:

for any correlation value $\rho \in \{0, 1/20, 2/20, \ldots, 1\}$, we have considered $N = 1000$ samples, with size $n = 50$, of centered bivariate Gaussian couples with marginal variances equal to 1 and covariance ρ varying from 0 to 1. Recall that the noncorrelation test of Pearson, for this particular Gaussian model, is the most uniformly powerful test, among all tests with the same level α. Figure 3 presents the power curves for KL-MI (plain black curve), χ^2-MI (dotted black curve) independence tests, and Pearson (dashed red curve), Kendall and Spearman (mixed dashed and dotted red and blue curves) correlation tests, obtained from $N = 1000$ samples of size $n = 50$ of bivariate Gaussian distributions. For this setting, we can see form Figure 3, that our poposed KL-MI independence test is almost as powerful as the most uniformly powerful independence test of Pearson. χ^2-MI, Spearman and Kendall tests have comparable powers, lower than KL-MI and Pearson's ones.

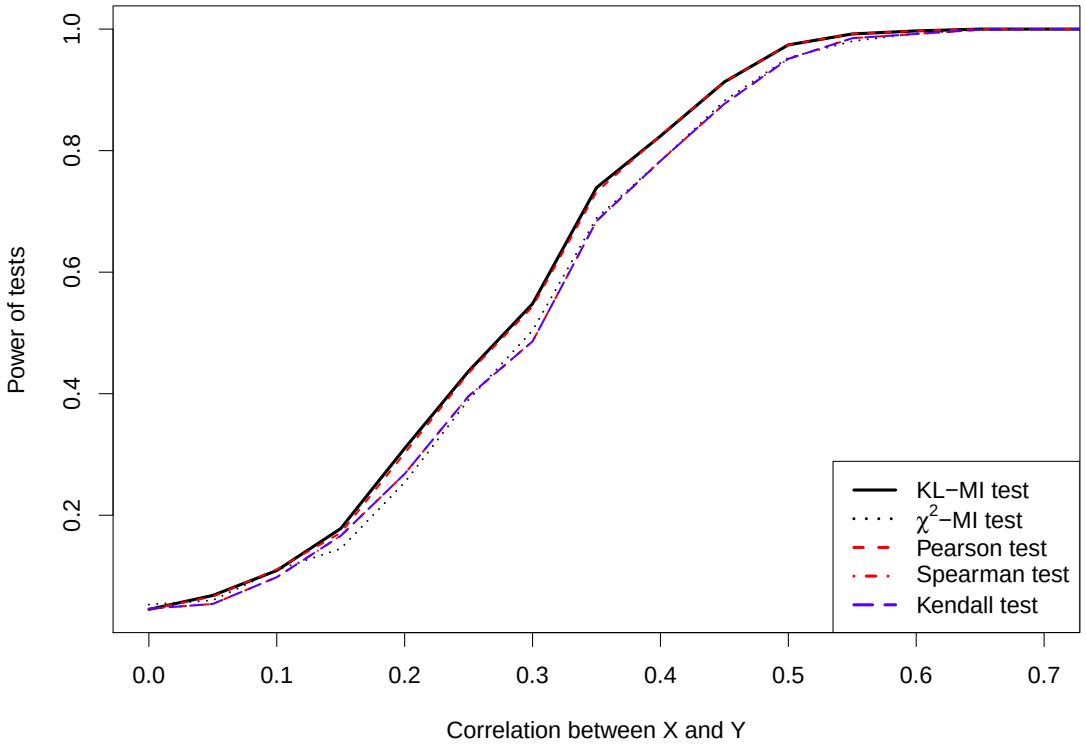

FIGURE 3. Comparison of powers of KL-MI and χ^2-MI tests with noncorrelation tests of Pearson, Spearman and Kendall.

5.3. **Comparison of φ-MI based tests for a copula density model.** This Section aims at comparing numerically the φ-MI based independence tests in the context of semiparametric copula-type model, as described in Example 2.5. We consider here the Farlie-Gumbel-Morgenstern (FGM) copula model

$$C_{FGM}(u,v;\theta) = uv(1+\theta(1-u)(1-v)), \quad (u,v)\in[0,1]^2, \quad \theta\in\Theta=[-1,1],$$

with $\theta_0 = 0$. We compare the powers of KL-MI and χ^2-MI based tests of independence to noncorrelation ones. We consider the alternative hypothesis that X and Y are uniformly distributed on $[0,1]$ and copulated by a FGM copula. We consider values of the parameter θ of the form $\theta = k/16$, with $k\in\{0,\ldots,16\}$. We have estimated the critical values of the KL-MI and χ^2-MI tests using the bootstrap procedure presented in Section 3.3, from an original sample of size $n = 50$ resampled 10 000 times. The powers are computed by Monte-Carlo method from $N = 5000$ samples of size $n = 50$. The results are presented in Table 3. We can see again that KL-MI based test still outperforms the others. We can see also that the nominal levels (of KL-MI and χ^2-MI test statistics) are sufficiently close to the test levels evaluated through the bootstrap procedure described in Section 3.3, with $\alpha = 0.05$.

θ	0	1/16	2/16	3/16	4/16	5/16	6/16	7/16
KL-MI	0.062	0.061	0.064	0.076	0.093	0.120	0.142	0.171
χ^2	0.054	0.055	0.057	0.066	0.084	0.108	0.129	0.160
Pearson	0.052	0.057	0.061	0.072	0.089	0.113	0.135	0.170
Spearman	0.055	0.058	0.060	0.069	0.086	0.110	0.133	0.164
Kendall	0.056	0.057	0.057	0.069	0.086	0.111	0.130	0.161

θ	8/16	9/16	10/16	11/16	12/16	13/16	14/16	15/16	1
KL-MI	0.219	0.261	0.312	0.382	0.431	0.498	0.565	0.622	0.691
χ^2	0.202	0.244	0.296	0.362	0.404	0.472	0.527	0.589	0.659
Pearson	0.213	0.257	0.309	0.375	0.427	0.493	0.549	0.611	0.677
Spearman	0.207	0.249	0.300	0.369	0.410	0.478	0.533	0.596	0.663
Kendall	0.203	0.243	0.293	0.356	0.405	0.467	0.527	0.584	0.647

TABLE 3. Power functions of KL-MI and χ^2-MI tests compared to noncorrelation tests obtained from $N = 5000$ samples of size $n = 50$ of the FGM copula with parameter θ varying from 0 to 1 by step of 1/16.

6. Concluding remarks and discussion

In this paper, we have defined and studied estimates of φ-mutual informations, based on the dual representation of φ-divergences and a semiparametric modeling of the density ratio between the joint distribution of the couple and the product distribution of its margins. The consistency

of these estimates – named dual-estimates – has been established assuming some classical regularity conditions on the model; the asymptotic normality has been established for classical Kullback-Leibler mutual information and specific models by means of classical M-estimation theory arguments. The asymptotic normality of other φ-mutual information dual-estimates may be derived similarly, for specific models depending on the considered φ-divergence. For example, when dealing with the power divergence associated to φ_γ functions given by (6), the asymptotic normality of the corresponding φ_γ-mutual-information dual-estimates may be derived in a similar way when focusing on the so-called γ-exponential semiparametric model

$$\mathbb{P} \in \left\{ P \in \mathcal{M}_1(\mathcal{X}\times\mathcal{Y}) \text{ such that } \frac{dP}{dP^\perp}(x,y) = \exp_\gamma\left(\sum_{k=0}^{d}\theta_k\xi_k(x)\zeta_k(y)\right), \theta=(\theta_0,\ldots,\theta_d)\in\Theta\right\},$$

where $\exp_\gamma(t) := ((\gamma-1)t+1)_+^{\frac{1}{\gamma-1}}$, with $(\cdot)_+ = \max(0,\cdot)$. Our semiparametric approach for estimating mutual informations constitutes a promising alternative to classical nonparametric procedures based on kernel density estimation or adaptive partitioning. No parameters such as bandwidth or kernel type has to be adjusted. The asymptotic normality of dual-estimates is also of significative importance, particularly, for hypothesis-testing purpose. For the sake of both completeness and accessibility, we are developing a package for the R software providing user-ready procedures, including the k-fold cross validation procedure described in Section 2.4, for selecting the model that best matches the data. We also aim at comparing the dual-estimates of mutual informations to nonparametric estimates. As an application of dual-estimation of mutual informations, we have derived a class of independence tests, recovering as a particular case, the classical χ^2-independence test. For a large variety of situations including finite-discrete random couples, the most efficient test is based on the KL-MI estimates, outperforming the classical χ^2-independence one. Motivated by the simulation experiments presented in this paper, we guess that the optimality of KL-MI independence test can be extended to a larger family of models.

7. Appendix

Proof of Proposition 3.1. Using continuity of $g_\theta(x,y)$ in θ on the compact set Θ, and condition (A.7), we can state, by Bienaymé-Tchebychev inequality, the uniform convergence in probability

$$A_n := \sup_{\theta\in\Theta}\left|\int g_\theta(x,y)\,\mathrm{d}\widehat{\mathbb{P}}^\perp(x,y) - \int g_\theta(x,y)\,\mathrm{d}\mathbb{P}^\perp(x,y)\right| \to 0. \tag{35}$$

Under condition (A.6), using continuity of $f_\theta(x,y)$ in θ over the compact set Θ, we have by uniform weak law of large numbers the convergence in probability

$$B_n := \sup_{\theta\in\Theta}\left|\int f_\theta(x,y)\,\mathrm{d}\widehat{\mathbb{P}}(x,y) - \int f_\theta(x,y)\,\mathrm{d}\mathbb{P}(x,y)\right| \to 0. \tag{36}$$

Now, we have

$$\begin{aligned}\left|\widehat{I}_\varphi - I_\varphi(\mathbb{P})\right| &= \left|\sup_{\theta\in\Theta} M_n(\theta) - \sup_{\theta\in\Theta} M(\theta)\right| \\ &= \left|M_n(\widehat{\theta}_\varphi) - M(\theta_T)\right| := |C_n| \end{aligned}\tag{37}$$

with

$$C_{n,L} := M_n(\theta_T) - M(\theta_T) \le C_n \le M_n(\widehat{\theta}_\varphi) - M(\widehat{\theta}_\varphi) =: C_{n,R}.$$

We can see that both sides converge in probability to zero, since

$$|C_{n,L}| \le A_n + B_n \quad \text{and} \quad |C_{n,R}| \le A_n + B_n$$

and the use of convergences (35) and (36). We conclude that $\widehat{I}_\varphi \to I_\varphi(\mathbb{P})$ in probability. The convergence of $\widehat{\theta}_\varphi$ to θ_T holds by direct application of Theorem 5.7 in van der Vaart (1998), using the uniform convergence in probability

$$\sup_{\theta\in\Theta}|M_n(\theta) - M(\theta)| \to 0$$

and the well-separability of the supremum θ_T; it is unique and interior point of Θ. □

Proof of Theorem 3.3. (a) Direct calculus gives

$$\mathbb{P}f'_{\mathbf{0}} - \mathbb{P}^\perp g'_{\mathbf{0}} = 0 \tag{38}$$

and

$$\mathbb{P}f''_{\mathbf{0}} - \mathbb{P}^\perp g''_{\mathbf{0}} = -\mathbb{P}^\perp\left(h'_{\mathbf{0}}h'^{\top}_{\mathbf{0}}\right) = -\Sigma_1. \tag{39}$$

Observe that the above matrix Σ_1 is symmetric and positive.
For any $\theta\in\Theta$, we have $M'_n(\theta) = \widehat{\mathbb{P}}f'_\theta - \widehat{\mathbb{P}}^\perp g'_\theta$. Note that

$$f'_{\mathbf{0}}(x,y) = g'_{\mathbf{0}}(x,y) = (1,\xi_1(x)\zeta_1(y),\ldots,\xi_d(x)\zeta_d(y))^\top.$$

We will state the asymptotic normality of $\sqrt{n}M'_n(\mathbf{0})$ using the multivariate Delta method. So consider the random column vector in $\mathbb{R}^{1+3d}$

$$V(X,Y) := (1,\xi_1(X),\ldots,\xi_d(X),\zeta_1(Y),\ldots,\zeta_d(Y),\xi_1(X)\zeta_1(Y),\ldots,\xi_d(X)\zeta_d(Y))^\top.$$

Denote by

$$\mu := \mathbb{E}(V(X,Y)) = (1,\mathbb{P}_1\xi_1,\ldots,\mathbb{P}_1\xi_d,\mathbb{P}_2\zeta_1,\ldots,\mathbb{P}_2\zeta_d,\mathbb{P}_1\xi_1\mathbb{P}_2\zeta_1,\ldots,\mathbb{P}_1\xi_d\mathbb{P}_2\zeta_d)^\top$$

which is a column vector in $\mathbb{R}^{1+3d}$. Then we have, by multivariate central limit theorem, the convergence in distribution

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} V(X_i,Y_i) - \mu\right) \to \mathcal{N}_{1+3d}\left(0,\Sigma\right),$$

with $\Sigma = \mathbb{E}\left((V(X,Y)-\mu)(V(X,Y)-\mu)^\top\right)$, from which we obtain, by multivariate Delta method,

$$\sqrt{n}\left(M_n'(\mathbf{0}) - \psi(\mu)\right) \to \mathcal{N}_{1+d}\left(\mathbf{0}, \Sigma_2 := \psi'(\mu)\Sigma\psi'(\mu)^\top\right), \tag{40}$$

where $\psi(\cdot)$ is the function defined on $\mathbb{R}^{1+3d}$ into $\mathbb{R}^{1+d}$ by

$$\psi(x_0,x_1,\ldots,x_d,y_1,\ldots,y_d,z_1,\ldots,z_d) = (0,x_1y_1-z_1,\ldots,x_dy_d-z_d)^\top$$

which is of class $\mathcal{C}^1$. Note that $\psi(\mu)=\mathbf{0}$, the first component of $M_n'(\mathbf{0})$ is equal to zero for all n and that the first column and row of the limiting covariance matrix Σ_2 are equal both to $\mathbf{0}$. Whence we have the convergence in distribution

$$\sqrt{n}M_n'(\mathbf{0}) \to \mathcal{N}_{1+d}\left(\mathbf{0},\Sigma_2\right). \tag{41}$$

By Taylor expansion of $U_n(\widehat{\theta}_{\varphi_1})$ in $\widehat{\theta}_{\varphi_1}$ around $\theta_T=\mathbf{0}$, using condition (A.8) and the convergence in probability of $\widehat{\theta}_{\varphi_1}$ to $\theta_T=\mathbf{0}$, we obtain

$$\mathbf{0} = M_n'(\widehat{\theta}_{\varphi_1}) = M_n'(\mathbf{0}) + M_n''(\mathbf{0})\widehat{\theta}_{\varphi_1} + o_{\mathbb{P}}(1)\widehat{\theta}_{\varphi_1}. \tag{42}$$

On the other hand, by (A.9), we can write

$$M_n''(\mathbf{0}) = \mathbb{P}f_{\mathbf{0}}'' - \mathbb{P}^{\perp}g_{\mathbf{0}}'' + o_{\mathbb{P}}(1) = -\Sigma_1 + o_{\mathbb{P}}(1).$$

Combining the last two displays, leads to

$$M_n'(\mathbf{0}) = \left(\Sigma_1 + o_{\mathbb{P}}(1)\right)\widehat{\theta}_{\varphi_1}. \tag{43}$$

We have, from (41), that $\sqrt{n}M_n'(\mathbf{0}) = O_{\mathbb{P}}(1)$, which by (43) implies that $\sqrt{n}\widehat{\theta}_{\varphi_1} = O_{\mathbb{P}}(1)$. Combining this last result with the relation (42), we obtain

$$\sqrt{n}\widehat{\theta}_{\varphi_1} = \Sigma_1^{-1}\sqrt{n}M_n'(\mathbf{0}) + o_{\mathbb{P}}(1). \tag{44}$$

Use this last relation and (41) to conclude the proof of part (a).

(b) By Taylor expansion of $\widehat{I}_{\varphi_1} = M_n(\widehat{\theta}_{\varphi_1})$, in $\widehat{\theta}_{\varphi_1}$ around $\theta_T=\mathbf{0}$, using the fact that $M_n(\mathbf{0})=0$ and some of the above statements, we obtain

$$\begin{aligned}\widehat{I}_{\varphi_1} &:= M_n(\widehat{\theta}_{\varphi_1}) \\ &= M_n'(\mathbf{0})\widehat{\theta}_{\varphi_1} - \frac{1}{2}\widehat{\theta}_{\varphi_1}^\top\Sigma_1\widehat{\theta}_{\varphi_1} + o_{\mathbb{P}}(1/n)\end{aligned}$$

which by (44) leads to

$$2n\,\widehat{I}_{\varphi_1} = \left(\sqrt{n}M_n'(\mathbf{0})\right)^\top \Sigma_1^{-1}\sqrt{n}M_n'(\mathbf{0}) + o_{\mathbb{P}}(1) \tag{45}$$

$$= \left(\sqrt{n}\Sigma_1^{-1/2}M_n'(\mathbf{0})\right)^\top \Sigma_1^{-1/2}\sqrt{n}M_n'(\mathbf{0}) + o_{\mathbb{P}}(1). \tag{46}$$

This proves the convergence in distribution of $2n\,\widehat{I}_{\varphi_1}$ to the random variable $Z^\top Z$, where Z is a centered multivariate normal random variable with covariance matrix $C = \Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}$. □

Proof of Theorem 4.1. First, under assumption (A.10), Eichelsbacher and Schmock (2002) yields the following large deviations principle for the sequence $(\widehat{\mathbb{P}})_n$ of empirical measures : we have for all measurable subset B of $\mathcal{M}_{\mathcal{G}}$,

$$\liminf_{n\to\infty}\frac{1}{n}\log\mathbb{P}^\perp\left(\widehat{\mathbb{P}}\in B\right) \geq -\inf_{Q\in\,\mathrm{Int}_{\tau_{\mathcal{G}}}(B)}\mathbb{K}(Q,\mathbb{P}^\perp), \tag{47}$$

and

$$\limsup_{n\to\infty}\frac{1}{n}\log\mathbb{P}^\perp\left(\widehat{\mathbb{P}}\in B\right) \leq -\inf_{Q\in\,\mathrm{Cl}_{\tau_{\mathcal{G}}}(B)}\mathbb{K}(Q,\mathbb{P}^\perp), \tag{48}$$

where $\mathrm{Int}_{\tau_{\mathcal{G}}}(B)$ and $\mathrm{Cl}_{\tau_{\mathcal{G}}}(B)$ denote, respectively, the interior and closure of B, with respect to the $\tau_{\mathcal{G}}$-topology. Since $Q\in\mathcal{M}_{\mathcal{G}}\mapsto\mathcal{D}_\varphi(Q,Q^\perp)$ is continuous, we obtain by contraction principle from (47) and (48), for all $d>0$,

$$\liminf_{n\to\infty}\frac{1}{n}\log\mathbb{P}^\perp\left(\widehat{I}_\varphi > d\right) \geq -\inf\left\{\mathbb{K}(Q,\mathbb{P}^\perp);\, Q\in\mathcal{M}_{\mathcal{G}} \text{ and } \mathcal{D}_\varphi(Q,Q^\perp)>d\right\} \tag{49}$$

and

$$\limsup_{n\to\infty}\frac{1}{n}\log\mathbb{P}^\perp\left(\widehat{I}_\varphi > d\right) \leq -\inf\left\{\mathbb{K}(Q,\mathbb{P}^\perp);\, Q\in\mathcal{M}_{\mathcal{G}} \text{ and } \mathcal{D}_\varphi(Q,Q^\perp)\geq d\right\}. \tag{50}$$

We now prove that the function $e_\varphi(\cdot) : d\in\mathbb{R}_+^*\mapsto\inf\{\mathbb{K}(Q,\mathbb{P}^\perp);\, Q\in\mathcal{M}_{\mathcal{G}} \text{ and } \mathcal{D}_\varphi(Q,Q^\perp)\geq d\}\in[0,+\infty]$ is right-continuous so that infima in (49) and (50) are equal, yielding (30). So, let $d>0$ be any positive real number, and show that $e_\varphi(\cdot)$ is right-continuous at d. If no $Q\in\Omega_d$ exists such that $\mathbb{K}(Q,\mathbb{P}^\perp)<+\infty$, obviously, since for any $d'\in\mathbb{R}_+^*$ such that $d\leq d'$, we have $\Omega_{d'}\subseteq\Omega_d$, then both $e_\varphi(d)$ and $e_\varphi(d')$ equal ∞, which implies that e_φ is right-continuous at d in this case. Now, assume that some $Q\in\Omega_d$ exists such that $\mathbb{K}(Q,\mathbb{P}^\perp)<\infty$. Two cases can be handled separately. First, assume that the infimum (31) is achieved for some Q such that $\mathcal{D}_\varphi(Q,Q^\perp)=:d'>d$. Then, for all d'' satisfying $d\leq d''\leq d'$, the equality $e_\varphi(d'')=e_\varphi(d)$ holds, yielding the right-continuity of e_φ at d. Second, assume that the infimum (31) is achieved for Q such that $\mathcal{D}_\varphi(Q,Q^\perp)=d$. Let us prove that there exists a sequence $(Q_n)_n$ of elements of $\{Q:\mathcal{D}_\varphi(Q,Q^\perp)>d\}$ such that $\mathbb{K}(Q_n,\mathbb{P}^\perp)\overset{n\to\infty}{\longrightarrow}\mathbb{K}(Q,\mathbb{P}^\perp)$ yielding right-continuity of $e_\varphi(\cdot)$ at

d. We build such a sequence $(Q_n)_n$ such that Q_n has the same marginal distributions as Q, i.e., $Q_{n,1} = Q_1$ and $Q_{n,2} = Q_2$. We have then $Q_n^\perp = Q^\perp$. Let

$$\overline{\theta} := \arg\sup_{\theta\in\Theta} \left\{ \int \varphi'(h_\theta)\, dQ - \int \varphi^*(\varphi'(h_\theta))\, dQ^\perp \right\},$$

so that

$$\int \varphi'(h_{\overline{\theta}})\, dQ - \int \varphi^* \left(\varphi'(h_{\overline{\theta}})\right)\, dQ^\perp = \mathcal{D}_\varphi(Q, Q^\perp) = d > 0. \tag{51}$$

Denote $\widetilde{Q}$ the image distribution on the Borel σ-field $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ of Q by the function $\varphi'(h_{\overline{\theta}})$. Let us prove by contradiction that $\widetilde{Q}$ can not be Dirac measure, by making use of either (A.12.a) or (A.12.b). If $\widetilde{Q}$ was a Dirac measure, necessarily $\varphi'(h_{\overline{\theta}})$ would be Q-a.s. constant, i.e., $h_{\overline{\theta}}$ would be Q-a.s. constant

$$h_{\overline{\theta}}(\cdot,\cdot) = c, \quad Q\text{-a.s.} \tag{52}$$

Now, if (A.12.a) holds, we can consider the set of all finite-discrete distributions with the same finite support $\mathcal{X} \times \mathcal{Y}$, instead of the set $\mathcal{M}_\mathcal{G}$. Hence, Q and $Q^\perp$ have same support, so that (52) implies that

$$h_{\overline{\theta}}(\cdot,\cdot) = c, \quad Q^\top\text{-a.s.} \tag{53}$$

Combining (52), (53) and (51), we obtain

$$\varphi(c) + \varphi'(c)(1-c) = d > 0. \tag{54}$$

On the other hand, by convexity of $\varphi(\cdot)$ and the fact that $\varphi(1) = 0$, we get

$$0 = \varphi(1) \geq \varphi(c) + \varphi'(c)(1-c) = d,$$

which contradicts the fact that $d > 0$. Alternatively, assume that (A.12.b) holds. Note that, under this assumption in connection with (A.2), we can see that the value θ_0 (of the parameter corresponding to independence) is necessarily $\theta_0 := (\alpha_0, \beta_0^\top)^\top = (0, 0, \ldots, 0)^\top$. We can see also, by contradiction as above, that $\overline{\theta}$ can not be of the form $(\overline{\alpha}, 0, \ldots, 0)^\top$ with $\overline{\alpha} \neq 0$. Hence, it can be written as

$$\overline{\theta} = (\overline{\alpha}, \overline{\beta}^\top)^\top \quad \text{with} \quad \overline{\beta} \neq (0, \ldots, 0)^\top. \tag{55}$$

Now, by (52), using the fact that $h_\theta(\cdot,\cdot)$ is of the form $\exp(\alpha + m_\beta(\cdot,\cdot))$, we get that

$$m_{\overline{\beta}}(\cdot,\cdot) = cte, \quad Q\text{-a.s.} \tag{56}$$

Note that the support of $Q^\perp$ is included in that of $\mathbb{P}^\perp$ (if not, Q would not be a.c.w.r.t. $\mathbb{P}^\perp$ and $\mathbb{K}(Q, \mathbb{P}^\perp)$ would not be finite). Hence, (56) implies that $\mathbb{P}^\perp(m_{\overline{\beta}}(X,Y) = cte) \neq 0$, which in turn implies that $\overline{\beta} = (0, \ldots, 0)^\top$ by assumption (A.12.b). This contradicts (55). We have proven then that $\widetilde{Q}$ is not a Dirac measure. So, there exist A, B two measurable subsets of

$(a_{\varphi^*}, b_{\varphi^*}) = \mathrm{Im}(\varphi')$ such that $\widetilde{Q}(A) > 0$, $\widetilde{Q}(B) > 0$ and $a := \inf A > b := \sup(B)$. Denoting $g := dQ/dQ^{\perp}$ the density of Q with respect to the product of its marginal distributions, set

$$\begin{aligned} g_n &:= \left(1+\frac{c_1}{n}\right) g\mathbb{1}_{\{\varphi'(h_{\overline{\theta}})\in A\}} + \left(1-\frac{c_2}{n}\right) g\mathbb{1}_{\{\varphi'(h_{\overline{\theta}})\in B\}} + g\mathbb{1}_{\{\varphi'(h_{\overline{\theta}})\in \overline{A\cup B}\}} \\ &= g + \frac{c_1}{n} g\mathbb{1}_{\{\varphi'(h_{\overline{\theta}})\in A\}} - \frac{c_2}{n} g\mathbb{1}_{\{\varphi'(h_{\overline{\theta}})\in B\}}, \end{aligned}$$

where $c_1 := \widetilde{Q}(B)$ and $c_2 := \widetilde{Q}(A)$. Note that g_n is nonnegative for n sufficiently large, and that $\int_{\mathcal{X}\times\mathcal{Y}} g_n(x,y)\, dQ^{\perp}(x,y) = 1$. Then, let Q_n be the probability distribution on $\mathcal{X}\times\mathcal{Y}$ such that $Q_{n,1} = Q_1$, $Q_{n,2} = Q_2$ and $\mathrm{d}Q_n/\mathrm{d}Q^{\perp} = g_n$. We have

$$\begin{aligned} \int_{\mathcal{X}\times\mathcal{Y}} \varphi'(h_{\overline{\theta}})\, dQ_n &= \int_{\mathcal{X}\times\mathcal{Y}} \varphi'(h_{\overline{\theta}}) g_n\, dQ^{\perp} \\ &= \int \varphi'(h_{\overline{\theta}})\, dQ + \frac{c_1}{n}\mathbb{E}_{\widetilde{Q}}(\mathrm{Id}.\mathbb{1}_A) - \frac{c_2}{n}\mathbb{E}_{\widetilde{Q}}(\mathrm{Id}.\mathbb{1}_B) \\ &\geq \int \varphi'(h_{\overline{\theta}})\, dQ + \frac{c_1}{n} a\widetilde{Q}(A) - \frac{c_2}{n} b\widetilde{Q}(B) \\ &= \int \varphi'(h_{\overline{\theta}})\, dQ + \frac{c_1 c_2}{n}(a-b) \\ &> \int \varphi'(h_{\overline{\theta}})\, dQ, \end{aligned}$$

where $\mathrm{Id}(x) := x$, for all $x \in (a_{\varphi^*}, b_{\varphi^*})$. Then,

$$\begin{aligned} \mathcal{D}_{\varphi}(Q_n, Q_n^{\perp}) &= \int \varphi'(h_{\overline{\theta}})\, dQ_n - \int \varphi^*(\varphi'(h_{\overline{\theta}}))\, dQ^{\perp} \\ &> \int \varphi'(h_{\overline{\theta}})\, dQ - \int \varphi^*(\varphi'(h_{\overline{\theta}}))\, dQ^{\perp} \\ &= d. \end{aligned}$$

Finally, the convergence of $\mathbb{K}(Q_n, \mathbb{P}^{\perp})$ to $\mathbb{K}(Q, \mathbb{P}^{\perp})$ can be proved using the decompositions

$$\mathbb{K}(Q_n, \mathbb{P}^{\perp}) = \mathbb{K}(Q_n, Q^{\perp}) + \mathbb{K}(Q^{\top}, \mathbb{P}^{\perp}), \quad \mathbb{K}(Q, \mathbb{P}^{\perp}) = \mathbb{K}(Q, Q^{\perp}) + \mathbb{K}(Q^{\top}, \mathbb{P}^{\perp}),$$

and Lebesgue's dominated convergence theorem. □

References

Beirlant, J., Dudewicz, E. J., Györfi, L., and van der Meulen, E. C. (1997). Nonparametric entropy estimation: an overview. *Int. J. Math. Stat. Sci.*, **6**(1), 17–39.

Broniatowski, M. and Keziou, A. (2006). Minimization of ϕ-divergences on sets of signed measures. *Studia Sci. Math. Hungar.; arXiv:1003.5457*, **43**(4), 403–442.

Cellucci, C. J., Albano, A. M., and Rapp, P. E. (2005). Statistical validation of mutual information calculations: Comparison of alternative numerical algorithms. *Phys. Rev. E*, **71**, 066208.

Chao, A. and Shen, T.-J. (2003). Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environ. Ecol. Stat.*, **10**(4), 429–443.

Cover, T. M. and Thomas, J. A. (2006). *Elements of information theory*. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition.

Darbellay, G. A. and Vajda, I. (1999). Estimation of the information by an adaptive partitioning of the observation space. *IEEE Trans. Inform. Theory*, **45**(4), 1315–1321.

Dudewicz, E. J. and van der Meulen, E. C. (1981). Entropy-based tests of uniformity. *J. Amer. Statist. Assoc.*, **76**(376), 967–974.

Eichelsbacher, P. and Schmock, U. (2002). Large deviations of U-empirical measures in strong topologies and applications. *Ann. Inst. H. Poincaré Probab. Statist.*, **38**(5), 779–797.

Genest, C., Ghoudi, K., and Rivest, L.-P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, **82**(3), 543–552.

Joe, H. (1997). *Multivariate models and dependence concepts*, volume 73 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Keziou, A. (2003). Dual representation of ϕ-divergences and applications. *C. R. Math. Acad. Sci. Paris*, **336**(10), 857–862.

Khan, S., Bandyopadhyay, S., Ganguly, A. R., Saigal, S., Erickson, D. J., Protopopescu, V., and Ostrouchov, G. (2007). Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data. *Phys. Rev. E*, **76**, 026209.

Konishi, S. and Kitagawa, G. (2008). *Information criteria and statistical modeling*. Springer Series in Statistics. Springer, New York.

Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating mutual information. *Phys. Rev. E*, **69**, 066138.

Liese, F. and Vajda, I. (1987). *Convex statistical distances*, volume 95 of *Teubner-Texte zur Mathematik [Teubner Texts in Mathematics]*. BSB B. G. Teubner Verlagsgesellschaft, Leipzig. With German, French and Russian summaries.

Moon, Y.-I., Rajagopalan, B., and Lall, U. (1995). Estimation of mutual information using kernel density estimators. *Phys. Rev. E*, **52**, 2318–2321.

Nelsen, R. B. (2006). *An introduction to copulas*. Springer Series in Statistics. Springer, New York, second edition.

Nikitin, Y. (1995). *Asymptotic efficiency of nonparametric tests*. Cambridge University Press, Cambridge.

Pardo, L. (2006). *Statistical inference based on divergence measures*, volume 185 of *Statistics: Textbooks and Monographs*. Chapman & Hall/CRC, Boca Raton, FL.

Shao, J. and Tu, D. S. (1995). *The jackknife and bootstrap*. Springer Series in Statistics. Springer-Verlag, New York.

Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Bünau, P., and Kawanabe, M. (2008). Direct importance estimation for covariate shift adaptation. *Ann. Inst. Statist. Math.*, **60**(4), 699–746.

Sugiyama, M., Suzuki, T., and Kanamori, T. (2012). *Density ratio estimation in machine learning*. Cambridge University Press, Cambridge. With a foreword by Thomas G. Dietterich.

Suzuki, T., Sugiyama, M., Sese, J., and Kanamori, T. (2008). Approximating mutual information by maximum likelihood density ratio estimation. *Proceedings of ECML-PKDD2008*, **4**, 5–20.

Tsybakov, A. B. and van der Meulen, E. C. (1996). Root-n consistent estimators of entropy for densities with unbounded support. *Scand. J. Statist.*, **23**(1), 75–83.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

Van Hulle, M. M. (2005). Edgeworth approximation of multivariate differential entropy. *Neural Computation*, **17**, 1903–1910.

Wang, Q., Kulkarni, S. R., and Verdú, S. (2005). Divergence estimation of continuous distributions based on data-dependent partitions. *IEEE Trans. Inform. Theory*, **51**(9), 3064–3074.

[1]Laboratoire de Mathématiques de Reims EA 4535 and ARC-Mathématiques CNRS 3399, Université de Reims Champagne-Ardenne, France
E-mail address: amor.keziou@univ-reims.fr

[2]Laboratoire de Mathématiques de Reims EA 4535 and ARC-Mathématiques CNRS 3399, Université de Reims Champagne-Ardenne, France
E-mail address: philippe.regnault@univ-reims.fr